Received: 15 September 2023 | Accepted: 13 November 2023

DOI: 10.1002/aaai.12148

HIGHLIGHT

Physical scene understanding

Jiajun Wu

Stanford University, Stanford, California, USA

Correspondence
Jiajun Wu, Stanford University, Stanford, CA, USA.
Email: jiajunwu@cs.stanford.edu

Funding information
Stanford University; National Science Foundation; Office of Naval Research; Air Force Office of Scientific Research; Massachusetts Institute of Technology

Abstract
Current AI systems still fail to match the flexibility, robustness, and generalizability of human intelligence: how even a young child can manipulate objects to achieve goals of their own invention or in cooperation, or can learn the essentials of a complex new task within minutes. We need AI with such embodied intelligence: transforming raw sensory inputs to rapidly build a rich understanding of the world for seeing, finding, and constructing things, achieving goals, and communicating with others. This problem of physical scene understanding is challenging because it requires a holistic interpretation of scenes, objects, and humans, including their geometry, physics, functionality, semantics, and modes of interaction, building upon studies across vision, learning, graphics, robotics, and AI. My research aims to address this problem by integrating bottom-up recognition models, deep networks, and inference algorithms with top-down structured graphical models, simulation engines, and probabilistic programs.

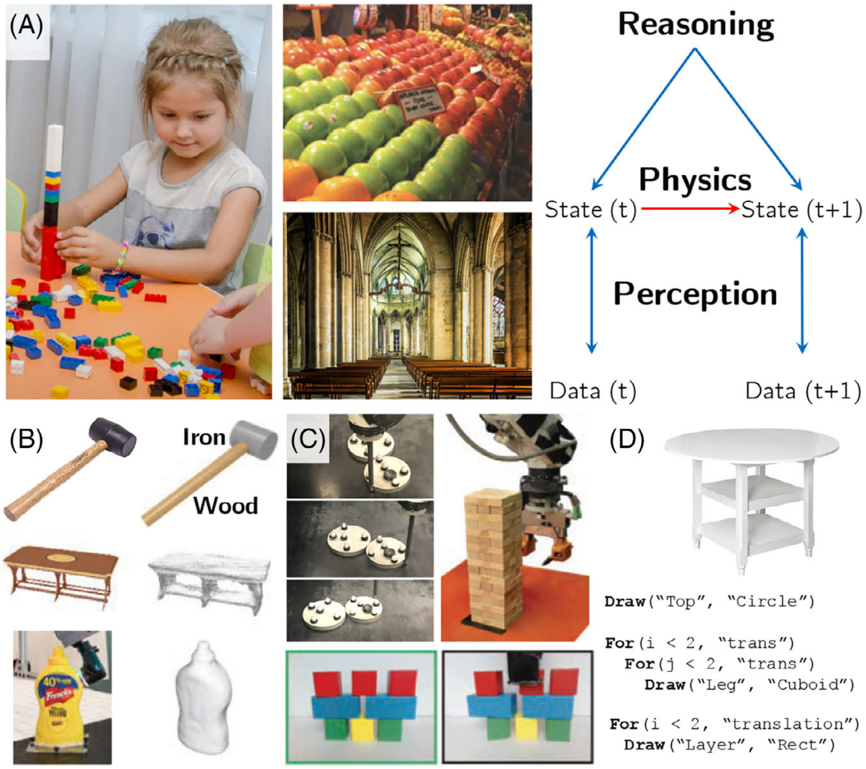
INTRODUCTION

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.

My research goal is to build machines that see, interact with, and reason about the physical world just like humans. This problem of **physical scene understanding** involves the following 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, multimodal sensory input (e.g., videos, sound, 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 build bridges from a single image and play challenging games such as Jenga. **Reasoning** (Figure 1D): How can physical models integrate structured, often symbolic, priors such as 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 interaction, 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



```
Draw("Top", "Circle")

For(i < 2, "trans")
  For(j < 2, "trans")
    Draw("Leg", "Cuboid")

For(i < 2, "translation")
  Draw("Layer", "Rect")
```

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

with top-down, structured graphical models, simulation engines, and probabilistic 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 can we build machines that have a human-like, physical understanding of complex, real-world 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.

Below I describe my research experience and future plans on the three research topics.

LEARNING TO PERCEIVE THE PHYSICAL WORLD

Motivated by human perception—rich, complex, generalizable, learning much from little—my research on perception has been centered on building structured, object-based models to characterize the appearance and physics of daily objects and scenes. These models integrate bottom-up deep recognition models with top-down simulation engines, and they learn by perceiving and explaining the physical world just like humans.

Seeing object intrinsics: shape, texture, and material. Drawing inspiration from human perception and computer graphics, my colleagues and I have built object appearance models that learn to perceive object intrinsics, such as shape, texture, and material, from raw visual observations, and to leverage such information for synthesizing new objects in 2D and 3D. The core object representation builds upon a coherent understanding of its intrinsic properties, in addition to extrinsic properties such as pose.

Our 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; Wu, Xue, et al. 2018) via modeling *intrinsic images*—depth, surface normals, and reflectance maps (Janner et al. 2017) (Figure 2A). Our research is inspired by the classic study 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 (Wu, Zhang, et al. 2018), to reconstructing object texture and material beyond geometry (Zhang et al. 2023), and to tackling cases where the object in the image is not from the training categories (Zhang et al. 2018).

Complementary to these bottom-up recognition models, we have also explored learning top-down graphics engines directly. We proposed 3D generative adversarial networks and point-voxel diffusion, among the first to apply generative-adversarial learning and diffusion to 3D shapes for unconditional shape synthesis (Wu, Zhang,

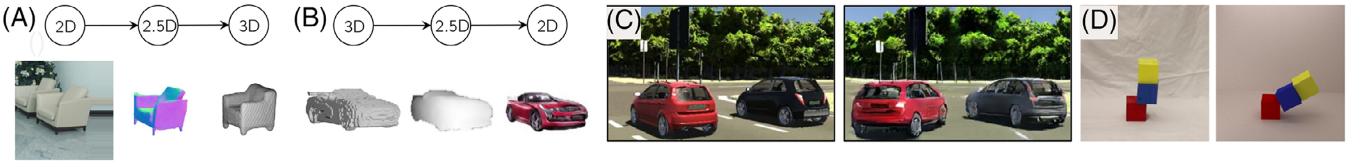


FIGURE 2 Learning to see shapes, texture, and physics. (A) Reconstructing 3D shapes from a single color image via 2.5D sketches (Wu, Wang, et al. 2017; Wu, Zhang, et al. 2018; Zhang et al. 2018; Janner et al. 2017). (B) Generative modeling of 3D shapes and 2D images via a disentangled representation for object geometry, viewpoint, and texture (Wu, Zhang, et al. 2016; Zhu et al. 2018; Chan et al. 2021; Zhou, Du, and Wu 2021; Zhang et al. 2023). (C) 3D-aware representations for objects and scenes (Wu, Tenenbaum, and Kohli 2017; Yao et al. 2018; Yu, Guibas, and Wu 2022; Yu, Agarwala, et al. 2023; Tian et al. 2023; Yu, Guo, et al. 2023). (D) Part-based object representations for its geometry and physics (Wu, Lim, et al. 2016; Wu, Lu, et al. 2017; Wu et al. 2015; Liu et al. 2018; Xu et al. 2019).

et al. 2016; Zhou, Du, and Wu 2021). These papers are influential; many other researchers have built on them. We have later extended the model as visual object networks (Zhu et al. 2018) and periodic implicit GANs (pi-GANs) (Chan et al. 2021), 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). We have generalized our models to scenes (Wu, Tenenbaum, and Kohli 2017; Yao et al. 2018; Yu, Guibas, and Wu 2022; Yu, Agarwala, et al. 2023), recovering structured scene representations that not only capture object shape and texture but enable 3D-aware scene manipulation (Figure 2C).

Seeing physics. Beyond object appearance, the intuition of object physics assists humans in scene understanding (Battaglia, Hamrick, and Tenenbaum 2013). We have developed computational models that learn to infer object physics directly from visual observations (Wu, Lim, et al. 2016; Wu et al. 2015). Our research on visual intuitive physics is the first in the computer vision community and has since led to many follow-up studies (Fragkiadaki et al. 2016; Mottaghi et al. 2016).

The Galileo model (Wu et al. 2015) 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 (Wu, Lim, et al. 2016), of 101 objects that interact in various physical events.

I have also worked on integrating geometry and physics perception (Figure 2D). For example, in visual de-animation (VDA) (Wu, Lu, et al. 2017), our model learns to jointly infer physical world states and simulate scene dynamics, integrating both a physics engine and a graphics engine. In physical primitive decomposition (PPD) (Liu et al. 2018), 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 dynamics-augmented neural objects (DANO) (Le Cleac'h et al. 2023), we enhance objects parametrized by neural implicit representations with their physical properties identified from raw observations; we then use such dynamic objects for future prediction. We have also extended these models to complex indoor scenes, exploiting stability for more accurate 3D scene parsing (Du et al. 2018).

Multimodal perception. Humans see, hear, and feel, perceiving the world through fusing multisensory signals. These signals play complementary roles: we see object shape and texture through vision, hear their material through sound, and feel their surface details through touch. In computer science, however, most recognition models and simulation engines primarily focus on visual data. Building upon techniques from the graphics community, we have been building generative audio–visual engines and using them for cross-modal perception (Zhang, Li, et al. 2017; Zhang, Wu, et al. 2017): how much do we know about objects from videos, and how much from audio? Our recent work includes developing a differentiable simulation model of impact sounds (Clarke et al. 2021) and building a benchmark for object impact sound fields (Clarke et al. 2023). Beyond auditory signals, we have also explored the integration of tactile signals with vision for better shape perception and reconstruction (Wang et al. 2018), and the integration of visual, auditory, and tactile information for robotic manipulation (Li et al. 2022).

In the past few years, we have also been developing a large-scale, multimodal, object-centric benchmark, ObjectFolder (Figure 3). It models the multisensory behaviors of both neural and real objects: it first includes 1000 neural objects in the form of implicit neural representations with simulated multisensory data (Gao et al. 2021, 2022); it also contains the multisensory measurements for 100 real-world household objects, based on a newly designed pipeline for collecting 3D meshes, videos, impact sounds, and tactile readings of real-world objects (Gao et al. 2023). ObjectFolder also has a standard benchmark suite of 10 tasks for multisensory object-centric learning, centered on object recognition, reconstruction, and manipulation

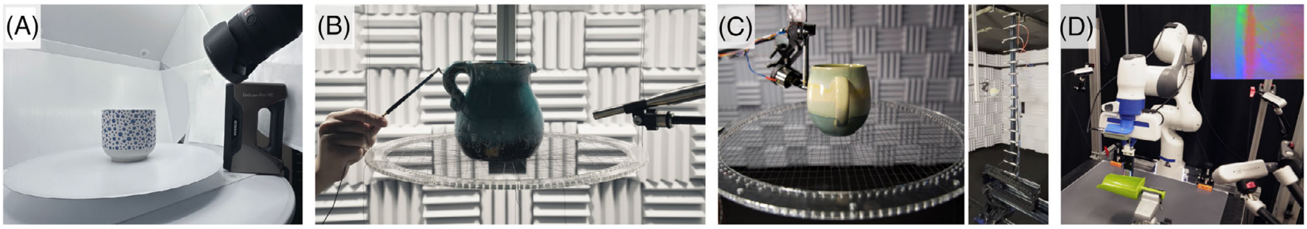


FIGURE 3 Multimodal perception (Gao et al. 2021, 2022, 2023; Clarke et al. 2021, 2023; Li et al. 2022). Visual (A): We use a scanner, a turntable, and a lightbox to acquire object geometry and texture. Auditory (B, C): We strike objects at precise points using an impact hammer, either by hand (B) or by a robot (C), recording the sound with a microphone array on a rotating gantry. The object is resting on a compliant mesh inside an acoustically treated room. Tactile (D): A robot presses a tactile sensor on the object with GelSight (Yuan, Dong, and Adelson 2017).

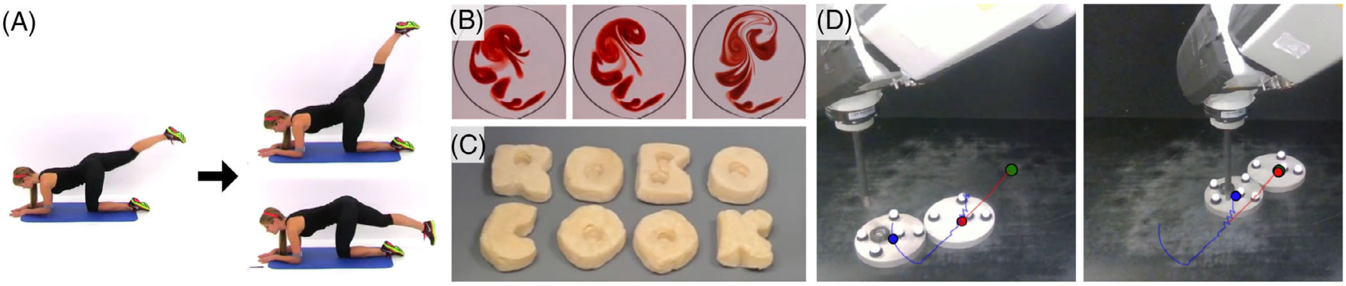


FIGURE 4 Physical models for future prediction and control. (A) Modeling visual dynamics allows us to generate multiple possible future frames from a single image (Xu et al. 2019; Xue et al. 2016). (B) Learned dynamics models support controlling soft robots Hu et al. (2019) and fluids Deng et al. (2023). (C) They also enable long-term manipulation of deformable objects and liquids (Li, Wu, Tedrake, et al. 2019; Shi et al. 2022, 2023). (D) 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, 2018).

with sight, sound, and touch. We have open-sourced both the datasets and the benchmark suite to catalyze and enable new research on multisensory object-centric learning in computer vision, robotics, and beyond (Gao et al. 2023).

PHYSICAL MODELS FOR REAL-WORLD INTERACTIONS

Beyond learning object-centric models from raw observations by inverting simulation engines, my research also includes learning to approximate simulation engines (forward models) themselves. Based on target domains and applications, my colleagues and I 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 4).

Compared with off-the-shelf simulators, a learned dynamics simulator flexibly adapts to novel environments and captures 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 4A). Our paper was among the first to consider uncertainty in the area of visual prediction. We have later extended the model to additionally capture the hierarchical structure among object parts (Xu et al. 2019).

Modeling dynamics directly in the pixel space is universal but challenging due to the entanglement of physics and graphics; an alternative is to separate perception from dynamics modeling and learn dynamics from object states. Our work along this line has shown that a model that learns to approximate object dynamics can be useful for planning (Janner et al. 2019), generalize to scenarios where only partial observations are available (Li, Wu, Zhu, et al. 2019), and discover physical object properties without supervision (Zheng et al. 2018; Le Cleac'h et al. 2023). We have further extended our model to particle-based representations so that it can characterize the dynamics of soft robots (Hu et al. 2019), fluids (Deng et al. 2023) (Figure 4B), and scenes with complex interactions among rigid bodies, deformable shapes, and liquids (Li et al. 2020; Li, Wu, Tedrake, et al. 2019; Shi et al. 2023) (Figure 4C).

We have also explored the idea of learning a hybrid dynamics model, augmenting analytical physics engines with neural dynamics models (Ajay et al. 2018) (Figure 4D). 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. The paper was selected as the Best Paper on Cognitive Robotics at the premier robotics conference (IROS 2018).

These dynamics models can be used in various control tasks: they help solve highly underactuated control problems (pushing disk A, which in turn pushes disk B to the target position) (Ajay et al. 2019), to control and co-design soft robots (Hu et al. 2019), to manipulate fluids and rigid bodies on a robot (Li, Wu, Tedrake, et al. 2019), to interact with plasticine to make complex shapes in multiple steps (Shi et al. 2022, 2023), and to interact and play games such as Jenga that involve complex frictional micro-interactions (Fazeli et al. 2018).

STRUCTURED PRIORS FOR COMMONSENSE REASONING

The physical world is rich but structured: natural objects and scenes are compositional (scenes are made of objects which, in turn, are made of parts); they often have program-like structures (objects are symmetric and made of evenly spaced repetitive parts). My colleagues and I have been exploring ways to bridge structured, often symbolic, priors into powerful deep recognition models. In previous sections, we have seen perception models that invert simulation engines and physical dynamics models that approximate simulation engines themselves. Here, we move one step further to learn the representation priors these simulation engines have—why they represent the world in the way they currently are.

A test of these neuro-symbolic representations is how well they support solving various reasoning tasks such as analogy making and question answering. Our work has demonstrated that when combined with deep visual perception modules, a symbolic reasoning system achieves impressive performance on visual reasoning benchmarks (Yi et al. 2018), 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) through curriculum learning (Mao et al. 2019), without human annotations.

Beyond static images, we have integrated neuro-symbolic representations with learned object-based dynamics models for temporal and causal reasoning on videos. On our newly proposed video reasoning benchmark, our model performs significantly better in answering all four types of questions: descriptive (e.g., “what color”), explanatory (“what’s responsible for”), predictive (“what will happen next”), and counterfactual (“what if”) (Yi et al. 2020; Chen et al. 2021). Similar ideas have been applied to visual grounding in 3D scenes (Hsu, Mao, and Wu 2023), human motion understanding (Endo et al. 2023), and robotic manipulation (Wang et al. 2023).

Learning symbolic structure is closely coupled with program synthesis. In particular, our recent work has made progress on the problem of inferring programs as a novel representation for shapes (Tian et al. 2019; Deng et al. 2022), scenes (Liu et al. 2019), and human motion (Kulal et al. 2021, 2022). This marks the start of our exploration of wiring highly structured, hierarchical priors into learning representations for physical scene understanding.

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) have begun to reintegrate. We should now take a more integrative view of these areas and actively explore their interactions for a more general AI landscape.

One such direction is to achieve a more fundamental integration of perception, reasoning, and planning. Although most computational models have treated them as disjoint modules, we observe that having them communicate with each other facilitates the model design and leads to better performance (Janner et al. 2019; Veerapaneni et al. 2019). For example, AI researchers have been integrating perception and planning in belief space (Kaelbling and Lozano-Pérez 2013)—our belief of the partially observable, uncertain world states. Building upon these insightful ideas, I would like to explore interactive perception by integrating both classic and modern AI tools: probabilistic inference for managing uncertainty; causal and counterfactual reasoning in generative models for explainability, imagination, and planning; and hierarchical inference for learning to learn, so knowledge builds progressively. In addition, discovering the cognitive and neural basis of perception, reasoning, and planning will be of significant value to understanding human intelligence.

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 real life, for which pure bottom-up, data-driven models cannot work well due to the exponentially increasing complexity. Our research has shown that they can learn to discover concepts and answer questions using only natural supervision (question–answer pairs) as humans (Mao et al. 2019; Yi et al. 2018; Hsu, Mao, and Wu 2023). In the future, I would like to explore the use of symbolic languages for knowledge representation and abstraction, and their integration with deep networks for flexible physical scene understanding and interaction.

Beyond physical objects and scenes, I want to build computational models that understand an agent's goals, beliefs, intentions, and theory of mind and use such knowledge for planning and problem-solving, drawing inspiration from intuitive psychology. While we have been inferring physical object properties from interactions, can we also build computational models that, just like 10-month-old infants (Liu et al. 2017), infer object values in agents' beliefs from their behaviors? Research in this direction would benefit the development of human-like and human-centered autonomous systems.

More generally, I want to connect computer science with other disciplines, such as cognitive science, neuroscience, social science, linguistics, and mechanical engineering. Research in cognitive science and neuroscience has been offering intuitions for AI researchers for decades; now, we are entering a new stage where contemporary research in intelligent systems or computer science, in general, may help us better understand human intelligence (Fischer et al. 2016; Yamins et al. 2014). Our research has suggested that computational models that combine bottom-up neural recognition networks and top-down simulation engines shed light on understanding cognitive and neural processes in the brain (Yildirim et al. 2019; Zhang et al. 2016). Much more work needs to be done in these areas. With the right integration of probabilistic inference methods, deep learning, and generative models, we can build more powerful computational models for both neural activities and cognitive, behavioral data. The same applies to developmental psychology. I want to compare and contrast human and artificial intelligence in understanding *core knowledge*—knowledge about object permanence, solidity, continuity, and containment, and concepts such as gravity and momentum (Spelke 2000). This interdisciplinary research deepens our understanding of multiple research areas and suggests future research topics.

We are in a unique and exciting time: the development of data, hardware, and algorithms (e.g., deep networks, graphical models, probabilistic programs) has enabled more flexible and expressive computational models. For the next decade, I believe building structured foundation models for machine physical scene understanding, as well as investigating its connection with perception, reasoning, and interaction, will be valuable and essential for developing computational systems that contribute to broad fundamental and practical research across disciplines.

CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict.

ORCID

Jiajun Wu https://orcid.org/0000-0002-4176-343X

REFERENCES

Ajay, Anurag, Maria Bauza, Jiajun Wu, Nima Fazeli, Joshua B. Tenenbaum, Alberto Rodriguez, and Leslie P Kaelbling. 2019. "Combining Physical Simulators and Object-Based Networks for Control." In *IEEE International Conference on Robotics and Automation (ICRA)*.

Ajay, Anurag, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P. Kaelbling, Joshua B. Tenenbaum, and Alberto Rodriguez. 2018. "Augmenting Physical Simulators with Stochastic Neural Networks: Case Study of Planar Pushing and Bouncing." In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Battaglia, Peter W., Jessica B. Hamrick, and Joshua B. Tenenbaum. 2013. "Simulation as an Engine of Physical Scene Understanding." *Proceedings of the National Academy of Sciences (PNAS)* 110(45): 18327–32.

Chan, Eric R., Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. "pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Zhenfang, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B. Tenenbaum, and Chuang Gan. 2021. "Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning." In *International Conference on Learning Representations (ICLR)*.

Clarke, Samuel, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L. James, and Jiajun Wu. 2023. "RealImpact: A Dataset of Impact Sound Fields for Real Objects." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Clarke, Samuel, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. 2021. "DiffImpact: Differentiable Rendering and Identification of Impact Sounds." In *Conference on Robot Learning (CoRL)*.

Deng, Boyang, Sumith Kulal, Zhengyang Dong, Congyue Deng, Yonglong Tian, and Jiajun Wu. 2022. "Unsupervised Learning of Shape Programs with Repeatable Implicit Parts." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Deng, Yitong, Hong-Xing Yu, Jiajun Wu, and Bo Zhu. 2023. "Learning Vortex Dynamics for Fluid Inference and Prediction." In *International Conference on Learning Representations (ICLR)*.

Du, Yilun, Zhijian Liu, Hector Basevi, Ales Leonardis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Learning to Exploit Stability for 3D Scene Parsing." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Endo, Mark, Joy Hsu, Jiaman Li, and Jiajun Wu. 2023. "Motion Question Answering Via Modular Motion Programs." In *International Conference on Machine Learning (ICML)*.

Fazeli, Nima, Miquel Oller, Jiajun Wu, Zheng Wu, Joshua B. Tenenbaum, and Alberto Rodriguez. 2018. "See, Feel, Act: Learning Complex Manipulation Skills with Causal Structure and Multi-Sensory Fusion." *Science Robotics* 4(26): eaav3123.

Fischer, Jason, John G. Mikhael, Joshua B. Tenenbaum, and Nancy Kanwisher. 2016. "Functional Neuroanatomy of Intuitive Physical Inference." *Proceedings of the National Academy of Sciences (PNAS)* 113(34): E5072–81.

Fragkiadaki, Katerina, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. 2016. "Learning Visual Predictive Models of Physics for Playing Billiards." In *International Conference on Learning Representations (ICLR)*.

Gao, Ruohan, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. 2021. "ObjectFolder: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations." In *Conference on Robot Learning (CoRL)*.

Gao, Ruohan, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. 2023. "The ObjectFolder Benchmark: Multisensory Learning with Neural and Real Objects." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gao, Ruohan, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. 2022. "ObjectFolder 2.0: A Multisensory Object Dataset for sim2real Transfer." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hsu, Joy, Jiayuan Mao, and Jiajun Wu. 2023. "NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, Yuanming, Jiancheng Liu, Andrew Spielberg, Joshua B. Tenenbaum, William T. Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. 2019. "ChainQueen: A Real-Time Differentiable Physical Simulator for Soft Robotics." In *IEEE International Conference on Robotics and Automation (ICRA)*.

Janner, Michael, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, and Jiajun Wu. 2019. "Reasoning about Physical Interactions with Object-Oriented Prediction and Planning." In *International Conference on Learning Representations (ICLR)*.

Janner, Michael, Jiajun Wu, Tejas D. Kulkarni, Ilker Yildirim, and Joshua B. Tenenbaum. 2017. "Self-Supervised Intrinsic Image Decomposition." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pack Kaelbling, Leslie, and Tomás Lozano-Pérez. 2013. "Integrated Task and Motion Planning in Belief Space." *The International Journal of Robotics Research (IJRR)* 32(9-10): 1194–227.

Kulal, Sumith, Jiayuan Mao, Alex Aiken, and Jiajun Wu. 2021. "Hierarchical Motion Understanding Via Motion Programs." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kulal, Sumith, Jiayuan Mao, Alex Aiken, and Jiajun Wu. 2022. "Programmatic Concept Learning for Human Motion Description and Synthesis." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Le Cleac'h, Simon, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. 2023. "Differentiable Physics Simulation of Dynamics-Augmented Neural Objects." *IEEE Robotics and Automation Letters (RA-L)*.

Li, Hao, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A. Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. 2022. "See, Hear, and Feel: Smart Sensory Fusion for Robotic Manipulation." In *Conference on Robot Learning (CoRL)*.

Li, Yunzhu, Toru Lin, Kexin Yi, Daniel Bear, Daniel L. K. Yamins, Jiajun Wu, Joshua B. Tenenbaum, and Antonio Torralba. 2020. "Visual Grounding of Learned Physical Models." In *International Conference on Machine Learning (ICML)*.

Li, Yunzhu, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. 2019. "Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids." In *International Conference on Learning Representations (ICLR)*.

Li, Yunzhu, Jiajun Wu, Jun-Yan Zhu, Joshua B. Tenenbaum, Antonio Torralba, and Russ Tedrake. 2019. "Propagation Networks for Model-Based Control Under Partial Observation." In *IEEE International Conference on Robotics and Automation (ICRA)*.

Liu, Shari, Tomer D. Ullman, Joshua B. Tenenbaum, and Elizabeth S. Spelke. 2017. "Ten-Month-Old Infants Infer the Value of Goals from the Costs of Actions." *Science* 358(6366): 1038–41.

Liu, Yunchao, Zheng Wu, Daniel Ritchie, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Learning to Describe Scenes with Programs." In *International Conference on Learning Representations (ICLR)*.

Liu, Zhijian, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Physical Primitive Decomposition." In *European Conference on Computer Vision (ECCV)*.

Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision." In *International Conference on Learning Representations (ICLR)*.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W. H. Freeman and Company.

Mottaghi, Roozbeh, Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. 2016. "Newtonian Scene Understanding: Unfolding the Dynamics of Objects in Static Images." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, Haochen, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. 2023. "RoboCook: Long-horizon Elasto-Plastic Object Manipulation with Diverse Tools." In *Conference on Robot Learning (CoRL)*.

Shi, Haochen, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. 2022. "RoboCraft: Learning to see, Simulate, and Shape Elasto-Plastic Objects with Graph Networks." In *Robotics: Science and Systems (RSS)*.

Spelke, Elizabeth S. 2000. "Core Knowledge." *American Psychologist* 55(11): 1233.

Tian, Stephen, Yancheng Cai, Hong-Xing Yu, Sergey Zakharov, Katherine Liu, Adrien Gaidon, Yunzhu Li, and Jiajun Wu. 2023. "Multi-Object Manipulation Via Object-Centric Neural Scattering Functions." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Yonglong, Andrew Luo, Xingyuan Sun, Kevin Ellis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Learning to Infer and Execute 3D Shape Programs." In *International Conference on Learning Representations (ICLR)*.

Veerapaneni, Rishi, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua B. Tenenbaum, and Sergey Levine. 2019. "Entity Abstraction in Visual Model-Based Reinforcement Learning." In *Conference on Robot Learning (CoRL)*.

Wang, Renhao, Jiayuan Mao, Joy Hsu, Hang Zhao, Jiajun Wu, and Yang Gao. 2023. "Programmatically Grounded, Compositionally Generalizable Robotic Manipulation." In *International Conference on Learning Representations (ICLR)*.

Wang, Shaoxiong, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T. Freeman, Joshua B. Tenenbaum, and Edward H. Adelson. 2018. "3D Shape Perception from Monocular Vision, Touch, and Shape Priors." In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Wu, Jiajun, Joseph J. Lim, Hongyi Zhang, Joshua B. Tenenbaum, and William T. Freeman. 2016. "Physics 101: Learning Physical Object Properties from Unlabeled Videos." In *British Machine Vision Conference (BMVC)*.

Wu, Jiajun, Erika Lu, Pushmeet Kohli, William T. Freeman, and Joshua B. Tenenbaum. 2017. "Learning to See Physics Via Visual De-Animation." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, Jiajun, Joshua B. Tenenbaum, and Pushmeet Kohli. 2017. "Neural Scene De-Rendering." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Jiajun, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T. Freeman, and Joshua B. Tenenbaum. 2017. "MarrNet: 3D Shape Reconstruction Via 2.5D Sketches." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, Jiajun, Tianfan Xue, Joseph J. Lim, Yuandong Tian, Joshua B. Tenenbaum, Antonio Torralba, and William T. Freeman. 2018. "3D Interpreter Networks for Viewer-Centered Wireframe Modeling." *International Journal of Computer Vision (IJCV)* 126(9): 1009–26.

Wu, Jiajun, Ilker Yildirim, Joseph J. Lim, William T. Freeman, and Joshua B. Tenenbaum. 2015. "Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, Jiajun, Chengkai Zhang, Tianfan Xue, William T. Freeman, and Joshua B. Tenenbaum. 2016. "Learning a Probabilistic Latent Space of Object Shapes Via 3D Generative-Adversarial Modeling." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, Jiajun, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T. Freeman, and Joshua B. Tenenbaum. 2018. "Learning Shape Priors for Single-View 3D Completion and Reconstruction." In *European Conference on Computer Vision (ECCV)*.

Xu, Zhenjia, Zhijian Liu, Sun Chen, Kevin Murphy, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "Modeling Parts, Structure, and System Dynamics Via Predictive Learning." In *International Conference on Learning Representations (ICLR)*.

Xue, Tianfan, Jiajun Wu, Katherine Bouman, and William T. Freeman. 2016. "Visual Dynamics: Probabilistic Future Frame Synthesis Via Cross Convolutional Networks." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex." *Proceedings of the National Academy of Sciences (PNAS)* 111(23): 8619–24.

Yao, Shunyu, Tzu-Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, William T. Freeman, and Joshua B. Tenenbaum. 2018. "3D-Aware Scene Manipulation Via Inverse Graphics." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yi, Kexin, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. "CLEVRER: Collision Events for Video Representation and Reasoning." In *International Conference on Learning Representations (ICLR)*.

Yi, Kexin, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yildirim, Ilker, Jiajun Wu, Nancy Kanwisher, and Joshua B Tenenbaum. 2019. "An Integrative Computational Architecture for Object-Driven Cortex." *Current Opinion in Neurobiology* 55: 73–81.

Yu, Hong-Xing, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, and Deqing Sun. 2023. "Accidental Light Probes." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, Hong-Xing, Leonidas J. Guibas, and Jiajun Wu. 2022. "Unsupervised Discovery of Object Radiance Fields." In *International Conference on Learning Representations (ICLR)*.

Yu, Hong-Xing, Michelle Guo, Alireza Fathi, Yen-Yu Chang, Eric Ryan Chan, Ruohan Gao, Thomas Funkhouser, and Jiajun Wu. 2023. "Learning Object-Centric Neural Scattering Functions for Free-Viewpoint Relighting and Scene Composition." *Transactions on Machine Learning Research (TMLR)*.

Yuan, Wenzhen, Siyuan Dong, and Edward H. Adelson. 2017. "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force." *Sensors*, 17(12): 2762.

Zhang, Renqiao, Jiajun Wu, Chengkai Zhang, William T. Freeman, and Joshua B. Tenenbaum. 2016. "A Comparative Evaluation of Approximate Probabilistic Simulation and Deep Neural Networks as Accounts of Human Physical Scene Understanding." In *Annual Meeting of the Cognitive Science Society (CogSci)*.

Zhang, Xiuming, Zhoutong Zhang, Chengkai Zhang, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. 2018. "Learning to Reconstruct Shapes from Unseen Categories." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Yunzhi, Shangzhe Wu, Noah Snavely, and Jiajun Wu. 2023. "Seeing a Rose in Five Thousand Ways." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Zhoutong, Qiujia Li, Zhengjia Huang, Jiajun Wu, Joshua B. Tenenbaum, and William T. Freeman. 2017. "Shape and Material from Sound." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, Zhoutong, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. 2017. "Generative Modeling of Audible Shapes for Object Perception." In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zheng, David, Vinson Luo, Jiajun Wu, and Joshua B. Tenenbaum. 2018. "Unsupervised Learning of Latent Physical Properties Using Perception-Prediction Networks." In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Zhou, Linqi, Yilun Du, and Jiajun Wu. 2021. "3D Shape Generation and Completion Through Point-Voxel Diffusion." In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhu, Jun-Yan, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Joshua B. Tenenbaum, and William T. Freeman. 2018. "Visual Object Networks: Image Generation with Disentangled 3D Representations." In *Advances in Neural Information Processing Systems (NeurIPS)*.

How to cite this article: Wu, J. 2024. "Physical scene understanding." *AI Magazine* 1–9. https://doi.org/10.1002/aaai.12148

AUTHOR BIOGRAPHY

Jiajun Wu is an Assistant Professor of Computer Science and, by courtesy, of Psychology 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 from the Massachusetts Institute of Technology. Wu's research has been recognized through the Young Investigator Programs (YIP) by ONR and by AFOSR, paper awards and finalists at ICCV, CVPR, SIGGRAPH Asia, CoRL, and IROS, dissertation awards from ACM, AAAI, and MIT, the 2020 Samsung AI Researcher of the Year, and faculty research awards from J.P. Morgan, Samsung, Amazon, and Meta.