Research Statement

Yunzhu Li

Humans have a strong intuitive understanding of their surrounding physical world (Figure 1). We observe and interact with the environment through multiple sensory modalities, including vision and touch. During the interactions, our brains process the sensory data, build a mental model of the world, and predict how the environment would change if we applied a specific action (i.e., intuitive physics). This predictive ability does not rely on analytical physics equations yet applies to objects of different materials (e.g., rigid bodies, deformable objects, and fluids), enabling a tremendous amount of interactive skills far superior to those of current robotic systems.

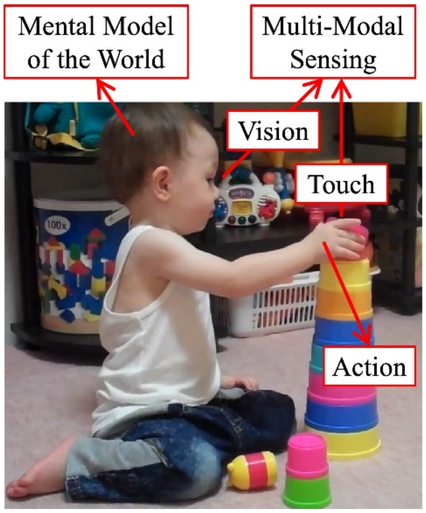
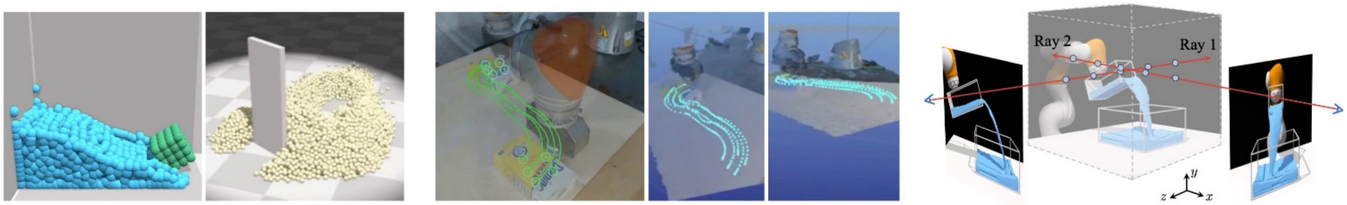


Figure 1: Humans perceive the world via multiple sensory modalities (e.g., vision and touch) and use the mental model in our brain to plan our actions.

My research goal is to develop robotic systems that learn from their physical interactions and build predictive models of the environment that can generalize widely across a diverse set of problems. To accomplish this kind of generalization, I have studied how to bring together the flexibility of deep neural network models with structured inductive biases, which enables broad generalization to a variety of tasks and environments. In doing so, my research has been among the first to take robotic learning from the manipulation of rigid objects to the manipulation of a variety of dynamic and flexible objects (e.g., fluids and deformable or granular materials), unlocking a fundamentally new set of capabilities. The research theme, **learning structured world models from and for physical interactions**, requires learning algorithms and infrastructures vastly different from traditional model-based pipelines and unstructured reinforcement learning systems. Significant challenges exist in building inductive biases at the correct level of abstraction across sensing, perception, dynamics modeling, and optimization that form a coherent system for effective training and real-world deployment. In my research, I tackle this challenge from the following three core angles:

- **Scene Representation and Dynamics Modeling**: Based on the sensory data, how should we represent the surrounding environment at the right level of abstraction? How can we characterize the structures of the underlying physical world, model the dynamics for objects of different materials, and achieve better generalization ability? How should we select the model class (e.g., linear models, convolutional neural networks, or graph neural networks), and how does it affect downstream model-based optimization?
- **Physical Inference and Model-Based Control**: Given the dynamics model, how can we exploit the underlying structures and formulate and solve the optimization problem for downstream tasks such as physical inference, i.e., state and parameter estimation, and model-based control, i.e., coming up with action sequences to achieve the desired target? And how can we use the solution in a closed loop and deploy it to real robots?
- **Multi-Modal Perception of Physical Interactions**: How can we develop multi-modal perception systems that capture the richness of human perception, e.g., vision and scalable/flexible tactile sensing, that provide detailed modeling of the physical interactions and that lay the foundation for physically grounded dynamics modeling?

My approach leverages advances in robotics, computer vision, and machine learning, **integrating structural priors into deep neural networks** to capture the structure of the physical system and provide better generalization ability outside the training distribution (Figure 2). For example, my research has been the first to combine particle-based scene representation with novel graph-structured neural networks for the dynamics modeling of complicated objects, such as fluids and deformable foam. Leveraging the structures embedded in the learned models, I constructed and solved the model-based planning problem that enables robots to **accomplish complicated manipulation tasks** (e.g., manipulating a pile of boxes, pouring a cup of water, and shaping deformable foam into a target configuration; Figure 3), which goes far beyond the capabilities of prior systems. We have also built dense tactile sensors in various forms (e.g., gloves, socks, vests, and robot sleeves), **constructed multi-modal sensing and learning platforms** to study human-environment interactions, and shown success in full-body pose estimation and hand-object dynamics modeling (Figure 4). These sensing platforms allow for more scalable and expressive modeling of the interaction process in the natural environment, which can inject the desired physical priors into our system and build more structured and physically grounded models of the world.



(a) Particle Representation for Fluids and Granular Materials

(b) Keypoint Representation for the Box's Dynamic Movements in 3D

(c) Neural Implicit Representation for Fluid Dynamics

Figure 2: **Scene Representation and Dynamics Modeling.** My research investigates representations of the environment at different levels, ranging from object-centric models to keypoints and particles, and has demonstrated successes in modeling objects of various materials, including rigid bodies, deformable objects, granular materials, and fluids.

Scene Representation and Dynamics Modeling

Unlike physics-based models, we humans are not constantly calculating analytical physics equations, such as $F = ma$, in our brains. Instead, we build an intuitive model of the environment that predicts the system's evolution purely from sensory observations. Such predictive ability, while intrinsic to humans and applicable to a wide variety of systems, remains challenging for state-of-the-art computational models. For example, representations commonly used in the literature, such as the 6-DoF pose and abstract latent vectors, cannot handle deformable and compositional scenes because they are not expressive enough and generalize very poorly. We need representations that can capture the complexity of these scenes and impose inductive biases grounded to the underlying system.

My research introduced models that transform high-dimensional sensory data into structural representations at different levels of abstraction and exploit the structure for dynamics modeling. The learned dynamics models are more sample efficient and allow the modeling of dynamic and flexible objects, such as fluids and deformable foams, with a performance that goes far beyond what was possible before.

Keypoints. In [1, 2], we proposed the use of object keypoints, which are learned in a self-supervised manner and tracked over time (Figure 2b). These keypoints anchor our model-based predictions, and through concrete experimental evidence, we showed that keypoints provide the following appealing properties: (i) the output is interpretable and in 3D space, allowing us to analyze the visual model separately from the predictive model, and (ii) they can apply to deformable objects and (iii) achieve category-level generalization.

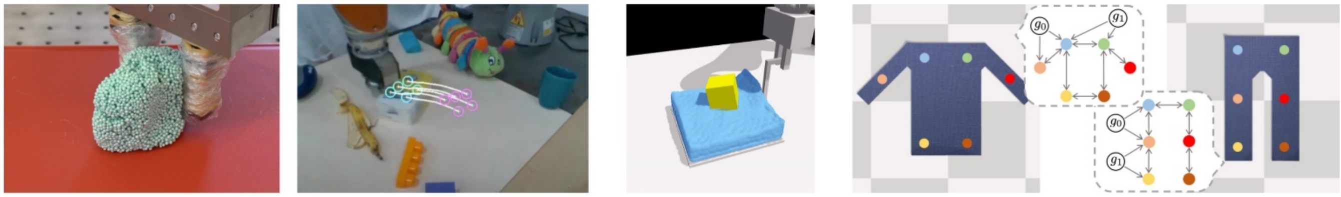
Particles. Keypoints are useful as a low-dimensional structural representation but cannot model objects with higher degrees of freedom, such as fluids. In [3, 4], we proposed dynamic particle interaction networks (DPI-Nets) to learn a particle-based simulator using graph neural networks (GNNs). Combining GNNs with particle-based systems acts broadly across objects of different materials and injects a strong inductive bias for learning: particles of the same type are governed by the same dynamics (Figure 2a). Such structural prior embedded in the model allows more effective training and better extrapolation generalization performance. We showed that our model can predict the dynamics of a wide variety of objects, including rigid bodies and challenging objects such as plasticine and fluids, and generalize to physical systems that are much larger than what the model was originally trained on.

Object-Centric and Hierarchical Representations. For tasks that operate on the object level, we have also investigated object-centric representations and graph-based dynamics models for neuro-symbolic reasoning [5]. We further extended it to hierarchical graph representations by mapping visual inputs to Physical Scene Graphs (PSGs) [6]. Within the learned graph, nodes represent objects or their parts, and edges represent relationships, in which many aspects of physical understanding become natural to encode, e.g., object permanence and shape constancy.

Latent and Implicit Representations. Prior works have shown impressive performance in learning world models in abstract latent space from visual observations yet are limited in generalization ability. Therefore, we incorporated the graph-based structural priors into the latent dynamics model and proposed propagation networks (PropNets) [8] that assume access only to partial observations and represent the system's state as a graph to capture the underlying compositionality to enable better generalization. Inspired by Koopman operator theory [9], we took a step further in [10] by enforcing linear constraints on the latent-space dynamics model in the PropNets for more efficient system identification and model-based optimization.

Besides the structures imposed on the dynamics model, we also seek to incorporate into our model inductive bias in the form of a learning-based differentiable renderer. In a recent paper selected for an **oral presentation** at a premier robot learning conference (i.e., CoRL) [11], we drew inspirations from advances in neural implicit representations [12]; we combined the differentiable volumetric renderer with an autoencoding framework and a latent dynamics model (Figure 2c). The resulting system greatly expands the model's capability and allows for (i) future prediction, (ii) out-of-distribution viewpoint generalization, and (iii) novel view synthesis in dynamic scenes involving complicated interactions between fluids and rigid objects.

Structured representations at different abstraction levels imply different modeling and generalization capabilities. Therefore, it is essential to understand their advantages and limitations, and in the future, I aspire to develop a *unified* framework that automatically selects the most suitable representations for a given task or adaptively changes the representation at different stages of a task.



(a) Deformable Object Manipulation

(b) Pusher-Slider in Visual Clutter

(c) Manipulating Fluid and a Floating Cube

(d) Causal Discovery in Physical Systems from Videos

Figure 3: **Physical Inference and Model-Based Control.** We tackled various downstream tasks by leveraging the structures in the learned dynamics models and powerful optimization tools from learning and control communities, and we demonstrated success in (d) physical scene reasoning and (a–c) manipulating objects made of different materials.

Physical Inference and Model-Based Control

In the previous section, we discussed how representations can be derived from sensory observations and used to learn the dynamics model, which can then be used for downstream inverse tasks, such as physical inference and model-based control. For example, we humans can estimate the weight of an object using the discrepancy between our mental model's prediction and the actual observation. We can also plan our behavior by imagining how our actions would change the environment and choosing the actions that produce the desired outcome. Although using forward models for inverse problems is not new and has been the subject of study in the robotics literature for a long time, it has proven challenging to find a general-purpose solution that is flexible and powerful enough to solve contact-rich and dynamic robotic manipulation tasks. In my work, I showed that with an appropriate choice of model, we can leverage the structures in the learned dynamics models and powerful insights gained from stochastic optimization to solve complex physical inference and model-based control problems from raw sensory observations.

Physical Inference. In [1], we considered the task of causal discovery from videos in an end-to-end fashion and without supervision on the ground-truth graph structure (Figure 3d). Our experiments show that, without retraining, the causal structure assumed by the model allows it to make counterfactual predictions and extrapolate to systems of unseen interaction graphs or graphs of different sizes from training. Beyond structural reasoning, we have also built a neuro-symbolic framework integrated with learned object-based dynamics models for temporal and causal reasoning about videos, which demonstrates significantly better performance in answering the following four types of questions: descriptive (e.g., "What shape?"), explanatory (e.g., "What is responsible for *x*?"), predictive (e.g., "What will happen next?"), and counterfactual (e.g., "What would have happened if _____ disappeared?") [5].

Model-Based Control. In [2, 11], we leveraged the parallel computing power of GPUs and applied sampling-based trajectory optimization methods to compute the control signals. Our model closes the control loop by taking the feedback from the environment and adjusting its planned behavior to compensate for the modeling error. We took a step further in [3, 8] by using the gradients from the learned model to achieve more efficient optimization of the action sequences. Through both simulated and real-world experiments, our robots have achieved success that greatly exceeds prior works in complex manipulation tasks, such as manipulating a deformable foam (Figure 3a), a pusher-slider system (Figure 3b), and a cup of water with floating ice cubes (Figure 3c). In [10], we traded off the expressiveness and efficiency of the model by assuming a linear structure over the latent state space, where we could apply quadratic programming to derive the control signals to manipulate ropes and control soft robots.

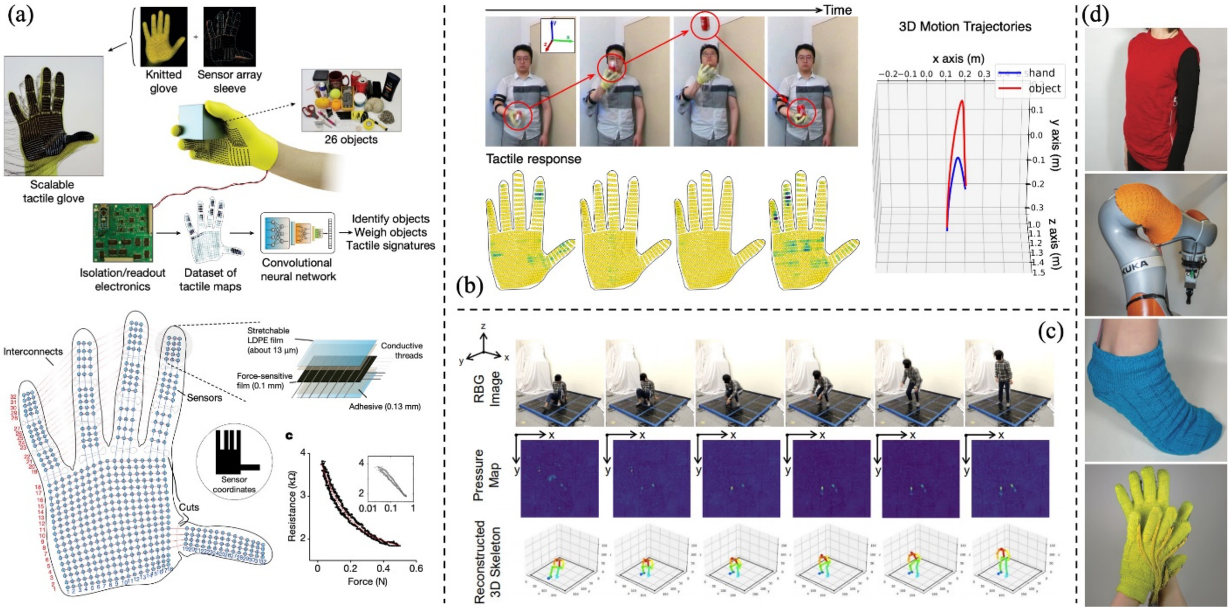


Figure 4: **Multi-Modal Sensing Platforms Using Vision and Touch.** To build better structured models of the world, we need expressive multi-modal sensing platforms that include the desired physical priors (e.g., contact and force constraints). Thus, we built scalable, deformable, and flexible tactile sensors in the form of gloves, socks, vests, and robot sleeves to obtain more detailed modeling of physical interactions and study the dynamic interactive behaviors of diverse physical activities. (a) A flexible glove capable of recording tactile information during hand-object interaction. (b) Tactile image and vision-enabled hand-object dynamics modeling. (c) Vision-supervised tactile learning for estimating 3D human poses. (d) 3D conformal tactile sensing garments for learning human-environment interactions.

Multi-Modal Perception of Physical Interactions Using Vision and Touch

To build structured dynamics models grounded in the underlying physical world, we require multi-modal sensing capabilities involving vision and touch to provide detailed contact information and inject physical constraints, such as contacting modes, force, and local contact patterns. Therefore, in papers accepted to *Nature* and *Nature Electronics*, we built multiple low-cost multi-modal sensing platforms to obtain synchronized data from vision and touch that constrain the models used for describing the physical interactions. The systems enable efficient learning frameworks of human-environment interactions, laying the foundation for more expressive structured dynamics modeling.

Specifically, we have developed a set of scalable, high-resolution, conformal tactile sensor arrays that can be automatically manufactured with inexpensive materials [13, 14]; for example, we have designed scalable tactile gloves (Figure 4a), which, when coupled with tools from the deep learning community, can learn to discover the signatures of human grasps [13]. Based on the same working mechanism, we have successfully developed various tactile sensing systems, including a large-scale intelligent carpet [15] (Figure 4c) and other wearable tactile sensor arrays (Figure 4d). The developed prototypes allowed us to record synchronized vision and tactile data from various human activities and human-object interactions [16], which can significantly facilitate the development of self-supervised learning systems and enable building models that capture the correlation between different sensory modalities and reflect constraints in the underlying physical world. For example, using the developed tactile glove, we have been able to demonstrate the ability to (i) identify objects, (ii) learn grasping patterns, (iii) identify weights of grasped objects [13], and (iv) model the dynamics of hand-object interactions by predicting the 3D locations of both the hand and the object purely from the tactile data (Figure 4b) [16]. The wearable garments further allow us to collect tactile information through human-environment interactions, enabling learning systems to identify 3D human poses, activities, and postures.

Insights from the multi-modal sensing platforms—through the lens of automatically manufactured dense tactile arrays—can aid in the future design of new prosthetics and robot grasping/interaction tools and enable more effective robotic manipulation and human-robot interactive capabilities by learning physically grounded predictive models for a more expressive and structured modeling of the interactive dynamics.

Future Research

In the future, robots should be able to operate in unstructured environments and perform complicated physical interactions as dexterously and effectively as we humans do. Over the past years, there has been impressive progress in enriching robots' capabilities, but their performance is still far from ideal for real-world practical deployments. Huge performance gaps exist in virtually every aspect of the system between robots and humans, ranging from perception and dynamics modeling to planning/control. Progress on this front requires interdisciplinary study involving various research areas, including but not limited to robotics, computer vision, machine learning, and control. The tools and techniques developed during the process would also deepen our understanding and expand the horizon of the respective subfields and suggest new research topics. Below, I identify three research directions that expand upon my past research and take steps towards tackling some of the open challenges in this area.

From Specialist to Generalist: Adaptive Scene Representations from Raw Sensory Inputs. Despite rapid progress over the past years, the automatic selection and adaptation of structured scene representations from high-dimensional observation data are still far from solved, especially for challenging objects, such as sushi, salad, and cloth, which undergo complex physical interactions with severe occlusions. However, humans' mental model of these objects, although not perfect, is still much better than the state-of-the-art computational models. For example, when buttoning a shirt, we do not require knowledge of the full state of the cloth and only have to focus on local regions, or when grabbing a mug, we have representations at different levels of abstraction at different stages of the operation. Replicating or even surpassing such capability is no easy task. It requires us to draw insights from the cognitive science and neuroscience communities and the development of new learning paradigms to integrate structural priors into the optimization procedure that allows automatic and adaptive representation selection for effective dynamics modeling. This problem, which we have termed online model order reduction, is intriguing and would be of great value in expanding robots' capabilities.

Model-Based Optimization Beyond Planning & Control. Prior works have shown impressive planning and control results through optimization using a learned dynamics model. However, a series of theoretical and practical questions remains unanswered, including questions about the role of learning and guarantees of robustness and optimality, and how to incorporate domain-specific knowledge. I wish to draw inspiration from and utilize the tools of the theoretical machine learning and control communities to make formal claims about the reliability of the derived controller. One idea would be to construct and plan within the trust regions calculated from the model's confidence in its prediction. Going beyond planning & control, the dynamics model is essentially a forward mapping from the inputs and model parameters to the system's outputs. I also aspire to use the model for other inverse problems, such as inverse design, the co-optimization of robot design and control, and material and drug discovery.

Toward the Metaverse: Seamless Transitions Between Simulated and Real-World Physical Interactions. Humans can hold an object in their hands and turn it over, feeling different parts and constructing a mental model of the object describing its shape, material properties, and dynamics. However, due to the inaccurate modeling of physical interactions, it has always been a big challenge for computational models to map the real world into a simulation or match a simulation to real-world observations. Our scalable multi-modal sensing platforms [13, 14] will provide more detailed and physically grounded 3D modeling of the scene and the interaction process from vision and touch; this will deepen our understanding of how different modalities complement each other and allow us to construct physics-based and/or learning-based simulators that can replicate the physical interactions and sensory responses, enabling seamless transitions between simulations and the real world. I envision the overall system can learn to model the geometry and characterize the scene dynamics from a few trials of interactions. It can then support interactive applications in virtual and augmented reality (VR/AR) and facilitate the sim-to-real and real-to-sim transfer of challenging robotic manipulation tasks, such as dexterous in-hand manipulation/reorientation, non-prehensile manipulation, and human-robot interactions.

Now is a particularly exciting time to work in this interdisciplinary area with the rapid development of (i) large-scale datasets and physics simulators, (ii) computation and robotic hardware, and (iii) algorithms. There is huge potential to build intelligent agents that can perceive and interact with the world with unprecedented performance. The Department of Computer Science at UIUC has an extremely strong group of students and researchers. I am excited to build close collaborations at the intersection of robotics, computer vision, and machine learning, while working more broadly with faculty across the fields of control theory, computer graphics, mechanical engineering, cognitive science, neuroscience, and computational fabrication.

References

[1] **Yunzhu Li**, Antonio Torralba, Animashree Anandkumar, Dieter Fox, and Animesh Garg. “Causal Discovery in Physical Systems from Videos.” In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Lucas Manuelli, **Yunzhu Li**, Pete Florence, and Russ Tedrake. “Keypoints into the Future: Self-Supervised Correspondence in Model-Based Reinforcement Learning.” In *Conference on Robot Learning (CoRL)*, 2020.

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

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

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

[6] Daniel M. Bear, Chaofei Fan, Damian Mrowca, **Yunzhu Li**, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B. Tenenbaum, Daniel L.K. Yamins. “Learning Physical Graph Representations from Visual Scenes.” In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[7] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. “Learning Latent Dynamics for Planning from Pixels.” In *International Conference on Machine Learning (ICML)*, 2019.

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

[9] B. O. Koopman. “Hamiltonian Systems and Transformations in Hilbert Space.” In *Proceedings of the National Academy of Sciences of the USA*, 1931.

[10] **Yunzhu Li**, Hao He, Jiajun Wu, Dina Katabi, and Antonio Torralba. “Learning Compositional Koopman Operators for Model-Based Control.” In *International Conference on Learning Representations (ICLR)*, 2020.

[11] **Yunzhu Li**, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. “3D Neural Scene Representations for Visuomotor Control.” In *Conference on Robot Learning (CoRL)*, 2021.

[12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.” In *European Conference on Computer Vision (ECCV)*, 2020.

[13] Subramanian Sundaram, Petr Kellnhofer, **Yunzhu Li**, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. “Learning the Signatures of the Human Grasp Using a Scalable Tactile Glove.” In *Nature* 569, 698–702 (2019).

[14] Yiyue Luo, **Yunzhu Li**, Pratyusha Sharma, Wan Shou, Kui Wu, Michael Foshey, Beichen Li, Tomas Palacios, Antonio Torralba, and Wojciech Matusik. “Learning Human-environment Interactions using Conformal Tactile Textiles.” In *Nature Electronics* 4, 193–201 (2021).

[15] Yiyue Luo, **Yunzhu Li**, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomas Palacios, Antonio Torralba, and Wojciech Matusik. “Intelligent Carpet: Inferring 3D Human Pose from Tactile Signals.” In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Qiang Zhang, **Yunzhu Li**, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B. Tenenbaum, Wojciech Matusik, and Antonio Torralba. “Dynamic Modeling of Hand-Object Interactions via Tactile Sensing.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2021.