ATTENTIVE ABSTRACTIONS FOR FLEXIBLE VISION-BASED ROBOT LEARNERS

Dinesh Jayaraman, Assistant Professor of Computer and Information Science, University of Pennsylvania
WEB:https://www.seas.upenn.edu/∼dineshj/, EMAIL:dineshj@seas.upenn.edu

General-purpose robots of the future will need learning to acquire skills to work with us in our homes, offices, farms, and hospitals, and they will need visual perception to operate in these dynamic, uncertainty-rife, open-world settings. However, vision-based robot learning today is inflexible and inefficient: it is overreliant on robot-and-task-specific training experiences, expert-engineered task specifications, and substantial computational resources. In comparison, animals and humans demonstrate far more nimble forms of sensorimotor learning[1;2]. What are our robots missing?

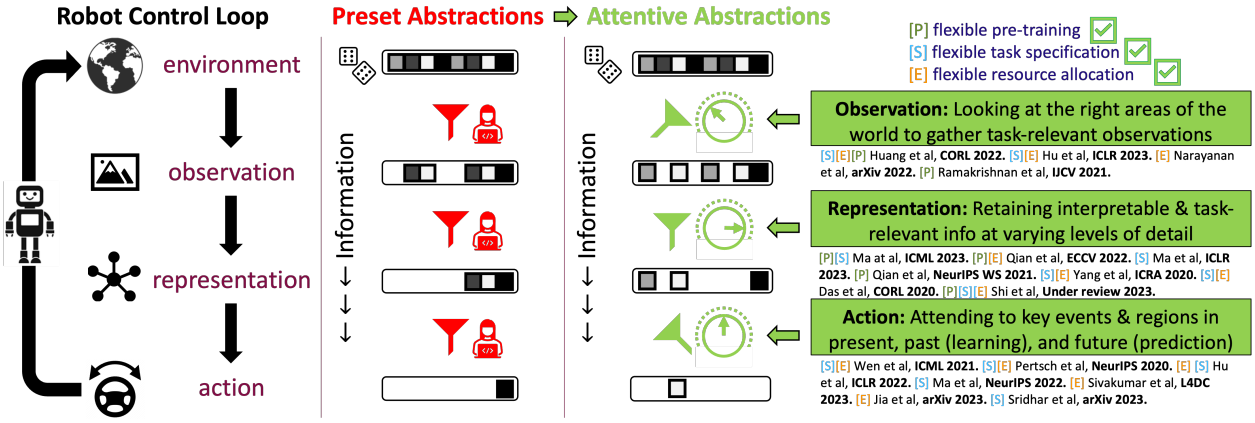

Fig. 1: In my research, "attention" steers information flow bottlenecks to retain and foreground task-relevant information in each module of a robot learner's control loop: observation, representation, decision making, and learning.

My research group's efforts target **flexible vision-based robot learning** for general-purpose robots, through **"attentive abstraction"** algorithms. Traditional modular robotic controllers contain inflexible, pre-set abstractions that bottleneck information flow between modules, hindering decision-making and learning. For example, a pre-defined state abstraction limits task capabilities. In trying to remove these bottlenecks, modern end-to-end approaches have gone too far: they produce large monolithic blackbox systems that are difficult to train and interrogate. In our attentive abstractions, "attention" is a mechanism that **steers the abstractions in a modular perception-action-learning loop to dynamically select task-relevant information**: which parts of the world to sense, how detailed of a state representation to use, which futures to predict, and which training data to learn from (Fig. 1). Thus, attentive abstractions overcome both the information-limitedness of pre-set modular controllers, and the sample-inefficiency of monolithic end-to-end learning systems. In doing so, they afford (Fig. 1, right): **(1) flexible pre-training** [P] of modules from diverse data sources, **(2) flexible task specification** [S] for robots to acquire new skills from layperson trainers, and **(3) flexible resource allocation** for efficiency [E] in compute, energy, and training data. In the rest of this document, I highlight these three types of flexibility with the tags [P] [S] [E].

I am excited about several key advances in vision-based robot learning that we have made pursuing this vision during my time at UPenn. We have efficiently trained robots to perform many kitchen tasks by mapping language task specifications to reward signals[3] **(ICML'23)** [P] [S], focused behavior cloning losses on discovered keyframes to overcome longstanding spurious correlate issues in robotic tasks such as autonomous driving[4] **(ICML'21)** [S] [E], discovered multi-level object-centric representations of robotic scenes for imitation learning of tabletop manipulation tasks from only a few tens of demonstrations[5;6] **(ECCV'22)** [P] [E] [S], and trained interactive behaviors for reward perception to autonomously guide robot policy learning for tasks like screwing and door locking[7] [S] [E] [P] **(CORL'22)**. Our work has been received well, with publications at the most selective top tier conferences across machine learning, robotics, and vision, and the **CORL'22 Best Paper Award**. I will now expand upon these and our other key contributions at each stage of the robotic control loop.

Task-Aligned Representations with Object-Centric Spatial Attention. What does it mean for a visual representation of a scene to be suitable for robotic control? We have been making progress on this problem on two fronts. First, building on my past work in self-supervised visual representation learning from egomotion[9] and from temporal continuity[10], we have constructed a new self-supervision objective that trains **representations as universal value functions**[3;8] **(ICML'23, ICLR'23)** [P] [S]: distances in representation space between the current image o and a goal g (specified as an image[8] or as a language phrase[3]) must match the goal-reaching value function $V^*(o, g)$: how good is the current state for reaching the goal? Intuitively, this **objective encourages the representation to capture task progress.** Two key properties of $V^*(o, g)$ enable **flexibility in training data sources**: first, the V value function does not take action arguments, and further, conditioning on task goal g means that we can train on data from many different tasks. Therefore, we pre-train our representations as value functions on large and diverse pre-recorded human video datasets[11;12] using a form of offline reinforcement learning. Our objective function can be expressed as a control-aware member of the contrastive learning family of self-supervision objectives. Empirically, our representations enable **imitation learning of task policies from minimal demonstrations** for various robot manipulation tasks, such as folding a towel. Further, since our representations encode task value, they permit backing out dense rewards for downstream reinforcement learning of new skills conditioned on image or language goals: this means that **in an unseen environment, on an unseen robot, having only ever seen human data, a simple picture or language phrase describing a desired outcome suffices** for the robot to teach itself a new skill.

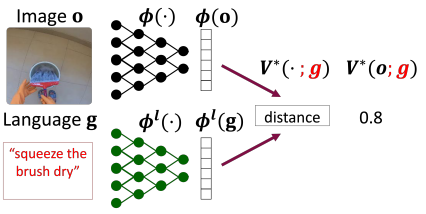

Fig. 2: Representations as value functions[3;8]

In parallel, we have been working to go beyond this standard scene vector representation format, towards representations formatted as **object-centric hierarchies**, mirroring the natural structure of the world. For example, a "pile of laundry" at finer resolutions could contain individual garments, their parts and keypoints, and eventually full 3D meshes. A hierarchical representation implicitly offers a coarse-to-fine menu of representations, so that, say, a laundry-folding robot could mix and match, **attending to different levels of abstraction** for different portions of the scene at each task phase. Building on my prior work on learning representations from object tracking[13], we first showed how to train flat object-centric image representations using optical flow signals extracted from training videos[14]. We have since shown that we can build "deformable keypoint pyramids" (DKP)[5] **(ECCV'22)** [P] [E]: hierarchies of landmark keypoints connected by springs between parent and children keypoints. We encourage coarse-to-fine gradation using a novel "assisted reconstruction" loss: broadly, each level of the keypoint pyramid must reconstruct the scene with assistance from an unstructured representation; and higher coarser levels are provided more assistance than lower, finer levels. On human and tabletop multi-object scenes generated by an exploring robot in our lab, DKP successfully discovers parent-children keypoint groups that are more consistent with manually annotated semantic keypoints than prior approaches. Once again, this permits scalable task specification: we have found that such keypoint-based object-centric representations enable **efficient policy learning from few demonstrations**[15] **or limited interactive experience**[16;17]. In ongoing research, we are exploring ways to distill such object-centric hierarchies straight out of large pre-trained models such as DINO-ViT[18] and SAM[19]. Further, using our V^*-based encoders from above to embed objects enables improved learning for various manipulation tasks on a real robot[6] (**Under review**, available on request) [P] [S] [E].

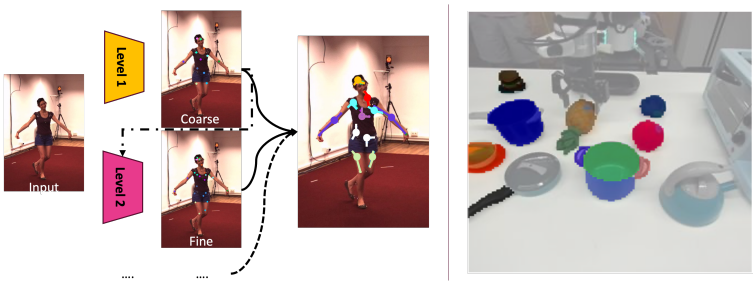

Fig. 3: (Left) Deformable keypoint pyramids[5]. **(Right)** Discovered object-based representations in our manipulation task setup[6]

Temporal Attention During Decision Making and Learning. How can an agent further selectively "attend" to such representations to select task-optimal actions? We have explored this question for visual model-based planning and offline policy learning. First, reliable predictive models can be used to select optimal actions, but are difficult to build for complex visual percepts over long time durations owing to compounding uncertainties. We have found one potential solution: "time-agnostic" jumpy predictors[20], which, given the freedom to select *which* future frame to attend to for prediction, consistently **select low-uncertainty "bottleneck" events that conveniently decompose long tasks into subtasks** to improve planning. Building on this, we have also proposed a hierarchical predictor[21] **(NeurIPS'20)** [S] [E] that first generates a coarse prediction of future waypoints, before iteratively filling in the missing frames at finer prediction levels, demonstrating large gains in long-horizon prediction and planning performance for manipulation tasks. Building further in this direction, we have successfully learned a "meta-controller" module[22] **(arXiv, Under review)** that selects how far out to predict, and how long to spend on plan optimization, **allocating time budgets to maximize success for a dynamic grasping task.**

This kind of **temporal attention is useful when applied not only to special *future* events during task execution, but also to special *past* events during learning.** We had identified a common "causal confusion" phenomenon[23–25], wherein imitators for tasks like autonomous driving discover shortcut solutions based on extrapolating past actions, rather than attend to the environment state. We have recently found a simple yet promising fix[4] **(ICML'21)** [S] [E]: we first train a decoy action extrapolation policy to mimic the expert data. Then, when training the real vision-based imitator, we boost weights for the samples where the decoy fails — environment observations most influence expert decisions at these "keyframes", such as when a traffic light turns green, or a nearby car brakes. A similar weighted behavior cloning objective also naturally emerges in our work on goal-conditioned offline reinforcement learning[26], where goal-reaching policies are trained to imitate appropriately weighted subsets of pre-recorded non-expert data, attending to the samples that most closely approach optimal behavior. Building on this work, we have recently developed a novel "policy-aware" model-based reinforcement learning[27] **(L4DC'23)** [E] technique that **focuses the training of a dynamics model on the most task-relevant transitions**, such that it can best inform faithful policy improvement within the learned model. Intuitively, a car driving on the road does not need to very precisely model the dynamics of driving on the rocks and can afford to downweight such experiences in its learning objective.

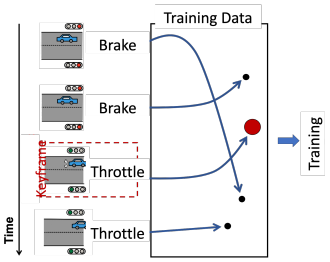

Fig. 4: Keyframe-focused visual imitation[4]

Finally, decision making also benefits from **attention over *spatial* regions.** We have recently proposed "robot-aware control"[28] **(ICLR'22)** [E] [S]: since robot geometry and kinematics are often known in advance for robot arms, we factorize visual models into an analytical robot predictor and a learned visual non-robot region predictor. This improves planning cost functions and dynamics model learning, and most importantly, facilitates easier training data transfer between robots: for example, we demonstrate "zero-shot" transfer of visual models from Franka to previously unseen WidowX robots to enable **skill transfer between robots**.

Attentive Information Gathering for Observation and Exploration. The observation and exploration stage acquires information generated in the environment into the robot's control loop for further processing into representations and decisions. "Attention" at this stage involves active and interactive perception: how should an agent select exploratory actions that reveal task-relevant environment information? Here, I have previously shown how reinforcement learning methods can produce active vision policies for scene and object category recognition in static scenes[29], and explored their transferability after training on unsupervised scene reconstruction objectives[30;31]. We have since extended this work to include *exploration*[32]: **dropped into an unknown environment,**

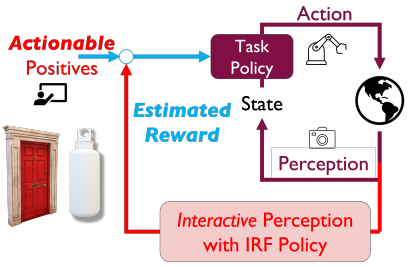

Fig. 5: "Interactive reward functions"[7] to train task policies.

how might a robot scope it out to be able to perform tasks in it afterwards? We have recently presented an exploration approach that **trains a goal-conditioned RL agent**[33] **(ICLR'23)** [S] [E]: it must simultaneously discover tasks, and also learn policies to accomplish them. Here, we show that the choice of goals we set at training time is critical: setting them to maximize the exploration value as predicted through a learned world model enables our approach to **discover skills like block stacking and accomplish them with *no supervision***.

Beyond exploration, we have also discovered a new use case for active/interactive perception in the robot learning setting. **Rather than estimating states as always, we use interactive perception to estimate *task rewards*** for training a robot with reinforcement learning[7] **(CORL'22)** [S] [E] [P]. Specifically, we train an interactive policy that inspects the outputs of the task policy as it learns, to determine what rewards to assign. This interactive policy in turn can be trained from examples of successfully completed tasks — it can learn how to distinguish a locked door from a merely closed one by tugging at the handle, or how to recognize a dust-free tabletop by running its fingers over it. Finally, these learned interactive verification behaviors can also be deployed at test time to improve task performance. This work enables **a new form of layperson-friendly task specification through physical object examples**, and received the **Best Paper Award at CORL 2022.**

Research Goals and Future Work. Fig. 1 presents these and our other relevant efforts during my time at UPenn, organized by the three stages of the control loop. As my research group matures, we hope to continue pushing the frontiers of flexible vision-based robot learning.

We plan to continue developing improved **robotics-ready visual representations** useful for manipulation and beyond. First, we are working towards a more cohesive unification of the two key streams of our research on representations: object-centric representations[5;6;14] as language-conditioned value functions[3;8], combining their complementary strengths. We will also incorporate sensing modalities beyond vision, and develop policy learning algorithms that exploit our structured object-centric hierarchies to dynamically select the minimal required representation at each instant for robust, sample-efficient learning and compute-efficient execution.

Second, I am interested in **safe and trustworthy learning**, a prerequisite for deploying learning robots in human environments. We have found in our early investigations that model-based reinforcement learning offers a promising route to safety. First, capturing uncertainty while learning a model can enable a simple and provably safe "pessimism under uncertainty" approach to safe exploration[34]. Further, pretraining on domain-randomized simulated environments can bootstrap such pessimistic exploration[35]. I hypothesize that the ultimate route to trustworthy machine learning systems must involve *shared representations* between humans and machines, such as our object-centric and language-grounded representations above.

Next, despite the attractiveness of **active and interactive visual perception**, robotics researchers have thus far found it difficult to exploit them for real robotic tasks due to training inefficiencies associated with a joint policy to make decisions about observation as well as task execution. We plan to explore solutions to this problem by factorizing the policy into individual modules which can each be trained largely independently of the other. This is true already of our interactive reward function policies[7] discussed above, but we are pursuing several directions generalizing this idea.

Fourth, underlying our investigations on attentive abstractions for robotic control are some **fundamental questions**. Can we characterize **what perceptual information is required for learning and executing optimal policies**, and how this requirement changes between tasks and task phases? Further, under resource constraints common in robotics such as compute, energy, or time, **how should we best utilize resources** across the different stages of the control loop? We have begun studying these questions. Control theoretical results impose fundamental limits on task performance for specific task families under partial observability (such as noisy visual perception or missing perceptual modalities); we have found that these limits are also predictive of empirical learning difficulties for reinforcement learning[36]. Our meta-controller results[22] above also show the advantages of dynamic time allocation in a time-constrained task.

Finally, we are exploring the newfound language abilities of autonomous systems to permit an interactive language-based interface between robot learners and human teachers. Now that we have already made initial progress on learning from various types of easy-to-provide task specifications such as demonstrations[4], image goals[8;28], language goals[8], and object goals[7], we are planning **an integrated system that can flexibly adapt to different teaching modes**. Just like it takes a village to raise a child, I eventually want to have small mobile robots exploring the GRASP lab at UPenn autonomously, and learning flexibly from any feedback or guidance that any interested student chooses to provide, in any fashion, without straitjacketing the supervision — they should be able to teach the robot as they might teach a pet or a child. This would be an important step towards the grand goal of placing learning robots in human homes, hospitals, farms etc., assisting the elderly, and aiding to automate dull, dangerous, and dirty jobs.

References

[1] John W Krakauer and Pietro Mazzoni. Human sensorimotor learning: adaptation, skill, and beyond. *Current opinion in neurobiology*, 21(4):636–644, 2011.

[2] Roland Sigrist, Georg Rauter, Robert Riener, and Peter Wolf. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. *Psychonomic bulletin & review*, 20:21–53, 2013.

[3] Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and **Dinesh Jayaraman**. LIV: Language-image representations and rewards for robotic control. *ICML*, 2023.

[4] Chuan Wen, Jierui Lin, Jianing Qian, Yang Gao, and **Dinesh Jayaraman**. Keyframe-focused visual imitation learning. *ICML*, 2021.

[5] Jianing Qian, Anastasios Panagopoulos, and **Dinesh Jayaraman**. Discovering deformable keypoint pyramids. *ECCV*, 2022.

[6] Junyao Shi, Jianing Qian, Yecheng Jason Ma, and Dinesh Jayaraman. Plug-and-play object-centric representations from "what" and "where" foundation models. *Under review*, 2023.

[7] Kun Huang, Edward Hu, and **Dinesh Jayaraman**. Training robots to evaluate robots: Example-based interactive reward functions for policy learning. *CORL (Best Paper Award)*, 2022.

[8] Yecheng Jason Ma, Shagun Sodhani, **Dinesh Jayaraman**, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via Value-Implicit Pre-Training. *ICLR (top 25 percent)*, 2023.

[9] **Dinesh Jayaraman** and Kristen Grauman. Learning image representations tied to ego-motion. *ICCV*, 2015.

[10] **Dinesh Jayaraman** and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. *CVPR*, 2016.

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (11):4125–4141, 2020.

[13] Ruohan Gao, **Dinesh Jayaraman**, and Kristen Grauman. Object-centric representation learning from unlabeled videos. *ACCV*, 2016.

[14] Jianing Qian and **Dinesh Jayaraman**. Object representations guided by optical flow. *NeurIPS 4th Robot Learning Workshop: Self-Supervised and Lifelong Learning*, 2021.

[15] Neha Das, Sarah Bechtle, Todor Davchev, **Dinesh Jayaraman**, Akshara Rai, and Franziska Meier. Model-based inverse reinforcement learning from visual demonstrations. *CORL*, 2020.

[16] Brian Yang*, **Dinesh Jayaraman***, Glen Berseth, Alexei Efros, and Sergey Levine. MAVRIC: Morphology-agnostic visual robotic control. *ICRA and IEEE RA-L*, 2020.

[17] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, **Dinesh Jayaraman**, and Roberto Calandra. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *ICRA and IEEE RA-L*, 2020.

[18] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[20] **Dinesh Jayaraman**, Frederik Ebert, Alexei A Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. *ICLR*, 2019.

[21] Karl Pertsch, Oleh Rybkin, Frederik Ebert, **Dinesh Jayaraman**, Chelsea Finn, and Sergey Levine. Long-horizon visual planning with goal-conditioned hierarchical predictors. *NeurIPS*, 2020.

[22] Yinsen Jia, Jingxi Xu, **Dinesh Jayaraman**, and Shuran Song. Learning a meta-controller for dynamic grasping. *arXiv preprint arXiv:2302.08463*, 2023.

[23] Pim de Haan, **Dinesh Jayaraman**, and Sergey Levine. Causal confusion in imitation learning. *NeurIPS*, 2019.

[24] Chuan Wen, Jierui Lin, Trevor Darrell, **Dinesh Jayaraman**, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. *NeurIPS*, 2020.

[25] Chuan Wen, Jianing Qian, Jierui Lin, Jiaye Teng, **Dinesh Jayaraman**, and Yang Gao. Fighting fire with fire: Avoiding dnn shortcuts through priming. *ICML*, 2022.

[26] Yecheng Jason Ma, Jason Yan, **Dinesh Jayaraman**, and Osbert Bastani. How far i'll go: Offline goal-conditioned reinforcement learning via f-advantage regression. *NeurIPS*, 2022.

[27] Yecheng Jason Ma, Kausik Sivakumar, Jason Yen, Osbert Bastani, and **Dinesh Jayaraman**. Learning policy-aware models for model-based reinforcement learning via transition occupancy matching. *L4DC*, 2023.

[28] Edward S. Hu, Kun Huang, Oleh Rybkin, and **Dinesh Jayaraman**. Know thyself: Transferable visuomotor control through robot-awareness. *ICLR*, 2022.

[29] **Dinesh Jayaraman** and Kristen Grauman. Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion. *ECCV*, 2016.

[30] **Dinesh Jayaraman** and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. *CVPR*, 2018.

[31] Santhosh K Ramakrishnan*, **Dinesh Jayaraman***, and Kristen Grauman. Emergence of exploratory look-around behaviors through active observation completion. *Science Robotics*, 2019.

[32] Santhosh K Ramakrishnan, **Dinesh Jayaraman**, and Kristen Grauman. An exploration of embodied visual exploration. *IJCV*, 2021.

[33] Edward Hu, Richard Chang, Oleh Rybkin, and **Dinesh Jayaraman**. Planning goals for exploration. *ICLR (top 25 percent) and CORL 2022 Robot Adaptation Workshop Best Paper Award*, 2023.

[34] Yecheng Jason Ma, **Dinesh Jayaraman**, and Osbert Bastani. Conservative offline distributional reinforcement learning. *NeurIPS*, 2021.

[35] Jesse Zhang, Brian Cheung, Chelsea Finn, Sergey Levine, and **Dinesh Jayaraman**. Cautious adaptation for reinforcement learning in safety-critical settings. *ICML*, 2020.

[36] Jingxi Xu, Bruce Lee, Nikolai Matni, and **Dinesh Jayaraman**. How are learned perception-based controllers impacted by the limits of robust control? *L4DC*, 2021.