Statement of Purpose

Humans skillfully parse visual streams to comprehend and interact with the world around us. My long-term research goal is to **endow robots with a visual system like ours**, so as to enable them to **adapt to unfamiliar scenarios** in highly unstructured environments.

For decades, robotic development was limited to industries that could afford the enormous hand-engineering efforts required. Recent advances, which instead *learn* behaviors from sensory inputs, have significantly eased this barrier to entry and expanded the scope of feasible robotic applications. However, when faced with unfamiliar scenarios, these systems are unable to extract meaningful structure from raw sensory observations, and hence struggle to make the correct conceptual links with their past experiences. To solve this *generalization* issue, a fundamental first step will be answering the question: *through what lens should robots view our rich and diverse visual world?*

I propose that our visual representation should be 1) abstract enough to endure changes in the environment, 2) expressive enough to support any new task, 3) optimized to respect data invariances, and 4) easily, or even autonomously, obtainable in any new environment. In my Ph.D., I hope to validate these hypotheses, propose benchmarks to quantify tradeoffs between these properties, and build algorithms to extract such representations. Ultimately, I imagine that once we can describe what a good representation looks like, vision-based robotics will have an "ImageNet moment." We will be able to design and collect a large-scale diverse dataset to enable open-world visual understanding for any robotics task.

Abstractness & Expressiveness. Robotic agents must correlate noisy raw observations with complex behaviors. However, high-dimensional visual inputs tend to contain distracting correlates that sidestep true reasoning. If we could amplify the relevant high-level signals and ignore these irrelevant low-level cues, would this improve reasoning, and hence generalization? As a first step toward this problem, I observed that conventional computer vision labels, such as object bounding boxes, are quite explicit in conveying high-level information. I devised a project[1] to evaluate how agents that view the world through high-level visual representations fare in unfamiliar environments, and presented it at ECCV '20.

I observed that our agent's performance was dependent on a delicate trade-off between the abstractness and expressiveness of its visual representation. It must be expressive enough for the agent to perform its task; however, if it is overly expressive, the agent may fail to generalize as it latches onto artifacts that do not exist in other environments or scenarios. Surprisingly, the strongest results used a hand-engineered representation rather than a standard computer vision label. This suggests that, rather than rely on conventional vision systems by default, we should explore learning-based approaches to find a representation that, as a first step, optimizes for the abstractness-expressiveness trade-off.

Invariance. Many prior works introduce effective data augmentations to boost supervisory signals and improve downstream performance. Some augmentations, such as trajectory-noise or DAgger-like methods, are fairly general-purpose, but others take advantage of task-specific invariances. For example, in a driving scenario, a birds-eye view representation can be freely rotated, along with its ground-truth steering, for extra expert demonstration training supervision. I argue that rather than rely on a delicate

[1] Brady Zhou*, Nimit Kalra*, Philipp Krähenbühl. *Domain Adaptation Through Task Distillation*. In ECCV 2020.

biasing of the training data, we should delegate this responsibility to the visual encoder. In doing so, downstream agents can easily latch onto the particular invariances that each downstream task obeys. At the same time, our world representation keeps an *open and flexible mind* by encoding many types of invariances, so as to generalize to novel tasks and unfamiliar scenarios as the agent experiences them.

To understand how encoding such an invariance into our visual representation would help, I set up a controlled experiment to compare 1) a contrastive objective between image-pose pairs with 2) a standard data augmentation trick (three cameras; Bojarski et. al., 2016) for avoiding lane drift. My preliminary experiments indicate that such a representation confers improved generalization to unseen driving scenarios. This suggests that we should view invariances as constraints to explicitly optimize, rather than hope our learning algorithm untangles them implicitly from complex behavioral supervision.

Supervision. Learning visual representations that optimize these properties may be difficult or impossible in the real world, as detailed state information is not available. However, we can often obtain task-specific supervision in the form of expert demonstrations. Can we use this readily available signal to learn a good visual representation that can later support novel tasks (i.e: isn't biased by the demonstrated task)? I was surprised to find that weak real-world signals, such as self-supervised dynamics objectives, act as a kind of regularization to encourage abstractness while maintaining expressiveness. In my early experiments, this enables a visual representation trained on a 'point navigation' task (PointGoal) to also solve a distinct 'exploration' task out-of-the-box. For more complex tasks, finding the right real-world signals will be very important, albeit difficult; however, my results suggest that weak signals may be a pathway forward.

Future Goals. Since the start of my junior year, I have been lucky to receive the guidance and mentorship of Professor Philipp Krähenbühl while working to understand and tackle these problems. In pursuing a Ph.D., I hope to spend my days collaborating with like-minded students on the problems of tomorrow. I aspire to become a professor in academia, and contribute back to the research community and my institution in the form of research insights, service, teaching, and mentorship.