Current computer vision methods are often formulated to map pixels to various output representations for different prediction tasks. Most research simply focuses on optimizing neural networks given fixed RGB image inputs. However, different types of inputs excel in different tasks. For example, Lidar depth sensors significantly improve color sensors for 3D recognition and power the perception of autonomous vehicles. Therefore why stop at prespecified inputs? Optimizing which sensory inputs are used by a network can improve performance in resource limited environments. Additionally, existing sensors are often limited and optimizing the parameters of a novel sensor (e.g. pixel spacing or underlying pixel filters) may reveal information that is not possible to acquire with current sensing technology. My research goal is to explore the collaboration of multimodal sensors, adaptive computational imaging algorithms, and downstream recognition models. My undergraduate experiences have gradually shaped this vision and built foundations for my career.

Research in 3D Perception. I have been working with Philipp Krähenbühl since my second year at UT Austin on the perception of autonomous vehicles. One open problem in Lidar-based 3D detection is the representation the model uses to reason about 3D objects. Most approaches represent objects as axis-aligned boxes in an overhead view to mimic the well-studied 2D detection problem. However, objects in 3D don't follow any particular orientation or size distribution. These box-based detectors have difficulties enumerating all orientations or fitting an axis-aligned box to rotated objects of various sizes and aspect ratios. I developed a new approach that represents objects as their center points. In contrast to boxes, center points are rotation and size invariant which better match the structure of 3D objects on the road**.** After detecting the centers, we can infer other box parameters more robustly. Experiments with various 3D backbones show that our center-based detector significantly outperforms box-based counterparts and achieves state-of-the-art performance on the two largest autonomous driving datasets (Waymo and nuScenes). ***I presented this work as first-author at CVPR 2021 and developed a corresponding codebase which has contributed to a community-wide adoption of our point-based representation for 3D detection, with over 100 citations within a year.*** This project gave me the confidence to lead future research and convinced me of the importance of simple and general representations in a world with constantly evolving sensory modalities and tasks.

In addition to developing better recognition models, identifying informative sensor inputs also provides gains for 3D recognition. For instance, while the Lidar sensor produces accurate depth measurements, they are often sparse at long range, making it hard to detect faraway objects as they easily fall between measurements. One solution to mitigate this challenge is multimodal fusion: Lidar and color sensors can complement each other. The combination produces higher-resolution measurements that contain richer semantic information useful for identifying foreground objects. The main challenge for camera-Lidar fusion is the domain gap between 2D images and 3D point clouds. To bridge the gap, I proposed to turn dense 2D pixels into virtual Lidar measurements through depth completion. These virtual points naturally integrate into any standard Lidar-based 3D detectors along with regular Lidar measurements. ***The final model, described in a first-author NeurIPS 2021 paper, significantly increases the detection accuracy for small and faraway objects.*** This project taught me the importance of utilizing multimodal information to overcome the limitations of separate sensors, much like how we as humans reason about the world using sensors like vision, hearing, and touch. A seamless combination of different senses will enable a more holistic understanding.

Research in MR Imaging. Since the summer after my sophomore year, I have been working with Katie Bouman and Yisong Yue from Caltech to develop methods for accelerated MRI. Accelerated MRI speeds up MRI acquisition by subsampling in the measurement space. The key challenge is to design the undersampling strategy to gather useful information from limited measurements due to time and cost constraints. Previous works often select all sampling locations in advance. However, MRI measurements are not taken all at once. We thus realized there existed a large potential in leveraging intermediate information to guide later sampling. Our approach enables adaptive sampling for individual patients to yield higher quality medical images under a fixed sampling budget. Experimental results show that our method consistently improves non-adaptive baselines for over 96% of test subjects. Besides sequential sampling, another source of improvement stems from our co-design of the learned sampling policy and image reconstruction model in an end-to-end framework. These findings strengthen my belief that a general computer vision system needs imaging algorithms that optimize the inputs and co-optimize with downstreamed task models. ***The resulting paper, which I co-first authored, was published at Machine Learning for Health 2021 and received the best paper award.***

In the future, I want to continue developing computer vision models that incorporate task structure, co-evolve with imaging sensors and algorithms, and yield practical impact. Moreover, I am interested in research directions across the full spectrum of the sensing-recognition collaboration, including multimodal supervision and interpretability.

Multimodal Supervision. Even without using a suite of sensors at test time, multimodal information can provide rich training signals. For example, monocular depth estimation uses a single camera for prediction but requires depth measurements for training. Additionally, it is often easier to annotate labels in one modality over the other for different tasks. Efficient reuse or transfer of supervision between different sensors may drastically reduce the labeling efforts and improve scalability. Just as we optimized the inputs in my previous MRI work, I am excited to explore how the supervision signal could also be optimized to reduce the burden on human labeling. My previous experience in multimodal object recognition and MRI sampling optimization will enable me to contribute to this field.

Interpretability. Many sensing and recognition algorithms are still uninterpretable. I aim to develop techniques that can reveal the underlying causal structure of a model's predictions, which is crucial for the deployment of safety-critical systems like autonomous vehicles. With the increasing use of computational sensors or multimodal fusion, understanding how different sensors work collectively or how computational cameras produce certain sampling patterns could provide valuable insights for developing efficient and effective vision systems. For example, such interpretable models may reveal an underlying pathology as the reason for an unusual optimized sampling pattern in our accelerated MRI framework.

One more paragraph about this specific school.