My research interests are in building AI models that are both practically useful and robust. Since beginning my undergrad at the **University of Washington**, I have concurrently been employed full-time as a researcher at the **Allen Institute for AI**. I have had the pleasure of being advised by **Ali Farhadi**, **Roozbeh Mottaghi**, and **Aniruddha Kembhavi**. I have also been fortunate to collaborate with **Richard Szeliski** on his computer vision textbook [8] and with **Ludwig Schmidt**. My work has led to 7 publications [1-7], including being the first-author on 5 of them. Most prominently, I initiated and led ProcTHOR [1], a recipient of the **Outstanding Paper Award at NeurIPS 2022**.

To make advances in AI, all modern approaches involve utilizing some dataset to train a model. Most research focuses on trying to build and optimize models to perform better on pre-defined tasks and datasets. However, the data available to the models fundamentally drive its capabilities. In computer vision, CLIP [9] and Stable Diffusion [10] work so well and robust because they leverage **internet-scale datasets** and **simple objective functions**. In NLP, Minerva [11] achieves incredible mathematical problem solving abilities by fine-tuning PaLM [12] on mathematical prompts, and InstructGPT [13] makes GPT-3 [14] notably more useful because it fine-tunes on prompts that resulted in particularly poor output from a model trained to perform next-token prediction on internet text. In my research, I focus on improving the **capabilities** and **robustness** of AI systems through **jointly optimizing training data and models**.

For many useful applications, massively diverse datasets are immensely **harder to obtain** than those that come from existing web images or text. For example, to build a general purpose household robot, it is clearly challenging to collect lots of trajectories in the real world. An emerging alternative to collecting data in the real world is to use **3D simulation**. AI2-THOR [6] is a pioneering 3D simulation framework for embodied AI research that allows researchers to generate large amounts of data quickly and easily. It includes highly realistic 3D household environments and objects, and has enabled research in many areas of AI, resulting in it being used for nearly 200 publications. As a core contributor, I find the aspect of being able to tweak the simulation engine to be able to collect any type of data incredibly powerful. The ability to easily create and build new tasks allows for much more flexibility in defining what capabilities we can endow in our models. Being able to write **efficient low-level 3D simulation code** enables me to research tasks like **visual room rearrangement** [7], scale the number of scenes from 200 to over 10K [1], and scale the number of objects from 2K to over 50K [2].

In computer vision, it is widely recognized that training on relatively small-scale datasets such as CIFAR-10 will not yield significant generalization. Yet, the mainstream approach in embodied AI was to hire artists to manually construct a small amount of 3D interiors that could be used for training. Across many tasks, training with such limited dataset inevitably resulted in **tremendous overfitting** to the training scenes and poor generalization in unseen test environments. A well-established approach to addressing overfitting is to scale the data to be more diverse. However, scaling data in embodied AI is incredibly difficult due to the time intensive process of having artists build more scenes. Thus, I proposed and led ProcTHOR [1], an ambitious project that aimed to use procedural **generative modeling for collecting training data**. Here, I independently built a generative function to sample massively diverse, realistic, and interactive 3D houses (Fig. 1). Through my initial experiments, I showed that training on such data leads to **remarkably robust generalization results**, which held true across a wide suite of tasks and environments.

Figure 1: ProcTHOR massively scales embodied AI by generating 3D houses that are used to train models.

While ProcTHOR enables remarkable generalization abilities, a clear goal is to **deploy models** trained in it to the real world. Here, successful models would have to (1) work robustly in *any* arbitrary real-world environment and (2) overcome the discrepancies between simulation to reality (*e.g.* physics, photorealism,

camera noise). In RoboTHOR [4], we rented an apartment in Seattle to study **sim-to-real transfer**. By training models in ProcTHOR and deploying them in the real-world, we noticed that simultaneously overcoming (1) and (2) was quite challenging. However, a key insight is that if we solved (1) by constructing a simulated reconstruction of the environment to more closely match its appearance (Fig. 2, left), then the models trained purely in simulation would perform quite well in reality.

Following the lessons of RoboTHOR, I developed Phone2Proc [3], a **conditional generative model** that samples simulated training scenes which semantically match a real-world target environment where a model will be deployed (Fig. 2, right). With just a **10-minute iPhone scan** of the environment, I built a model to generate a diverse dataset of scenes that semantically match the scanned environment. The generated scenes are based on the layout and arrangement of large objects in the scan, while small objects, lighting, materials, and clutter are randomized. Using Phone2Proc with a **simple RGB camera** improves the success rate of sim-to-real ObjectNav performance from 34.7% to 70.7% in five diverse environments. The model also makes agents robust to changes in the real world, such as with humans walking, rearranged objects, or added clutter. Conceptually, Phone2Proc enables the emergence of agents that can imagine multiple possible trajectories in their test environment, much like the **rollouts** used in the AlphaGo [15] algorithm. **Reimagining the training pipeline** to optimize the training dataset with learned knowledge about the test environment is an exciting area of research to me.

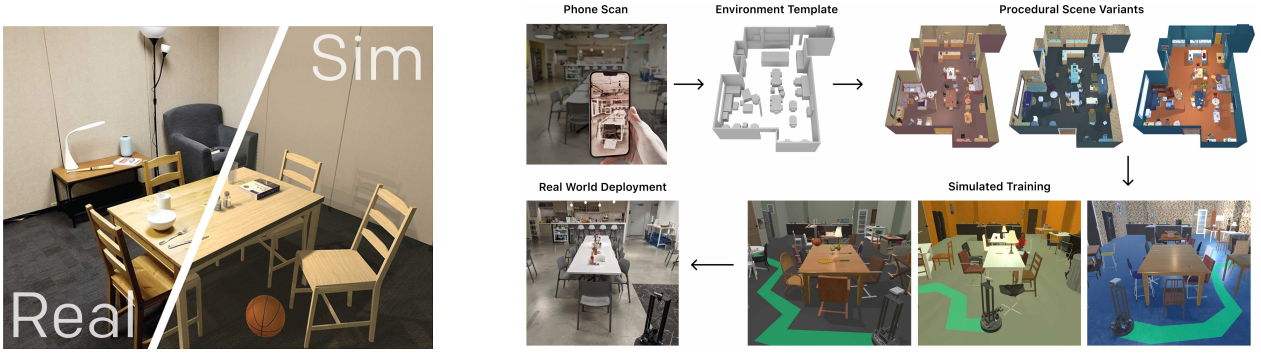

Figure 2: RoboTHOR (left) and Phone2Proc (right) focus on robustly solving sim-to-real distributional shifts.

Finally, while we live in a 3D world, most computer vision research happens on 2D images, in part due to the scarcity of high-quality 3D data (Fig. 3). Therefore, I led and initiated the development of Objaverse [2], an **annotated internet-scale 3D object dataset**. It includes over 800K visually diverse 3D objects, coverage on over 20K concepts, and includes text annotations and animations. By comparison, the next largest similar dataset is ShapeNet [16], which includes 50K visually similar 3D objects from 55 categories. In our initial experiments, we show the usefulness of Objaverse for **3D generative modeling**, as augmentation to achieve SoTA on **LVIS long-tail instance segmentation** [17], to enable **open vocabulary embodied AI**, and to study the rotational **robustness** of open vocabulary CLIP classification models. I am thrilled by the potential of Objaverse to advance the fields of computer vision, graphics, and robotics, and to open up many avenues of research in open vocabulary 3D computer vision.

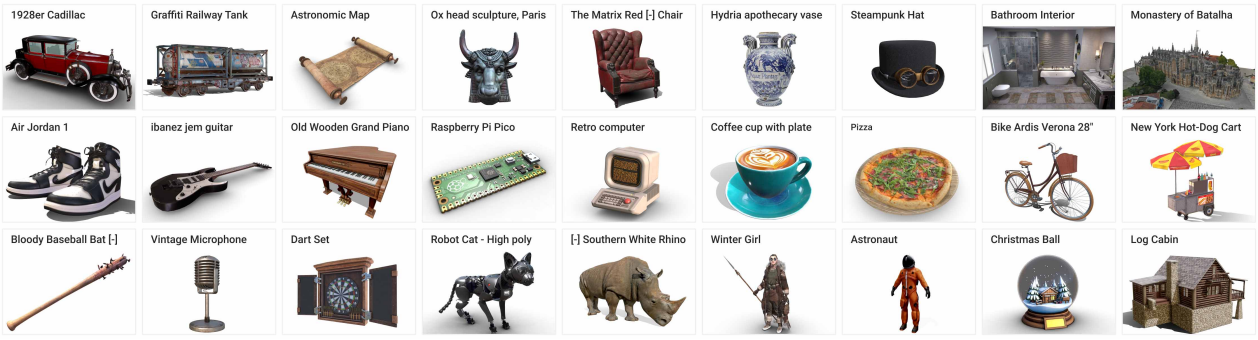

Figure 3: Objaverse is a massive annotated internet-scale 3D object dataset.

References

[1] **Matt Deitke**, Eli VanderBilt, Alvaro Hasteri, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, Roozbeh Mottaghi. *ProcTHOR: Large-Scale Embodied AI Using Procedural Generation.*
Outstanding Paper Award at NeurIPS 2022.

[2] **Matt Deitke**, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, Ali Farhadi. *Objaverse: A Universe of Annotated 3D Objects.* Under Review.

[3] **Matt Deitke**, Rose Hendrix, Luca Weihs, Ali Farhadi, Kiana Ehsani, Aniruddha Kembhavi. *Phone2Proc: Bringing Robust Robots Into Our Chaotic World.* Under Review.

[4] **Matt Deitke**, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, Ali Farhadi. *RoboTHOR: An Open Simulation-to-Real Embodied AI Platform.* CVPR 2020.

[5] **Matt Deitke**, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, Jiajun Wu. *Retrospectives on the Embodied AI Workshop.* ArXiv 2022.

[6] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, **Matt Deitke**, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, Ali Farhadi. *AI2-THOR: An Interactive 3D Environment for Visual AI.* ArXiv 2017.

[7] Luca Weihs, **Matt Deitke**, Aniruddha Kembhavi, Roozbeh Mottaghi. *Visual Room Rearrangement.* CVPR 2021 (Oral Presentation).

[8] Richard Szeliski. *Computer vision: algorithms and applications.* Springer Nature, 2022.

[9] Alec Radford, *et al. Learning transferable visual models from natural language supervision.* PMLR 2021.

[10] Robin Rombach, *et al. High-resolution image synthesis with latent diffusion models.* CVPR 2021.

[11] Aitor Lewkowycz, *et al. Solving Quantitative Reasoning Problems with Language Models.* NeurIPS 2022.

[12] Aakanksha Chowdhery, *et al. PaLM: Scaling Language Modeling with Pathways.* ArXiv 2022.

[13] Long Ouyang, *et al. Training language models to follow instructions with human feedback.* NeurIPS 2022.

[14] Tom Brown, *et al. Large models are few-shot learners.* NeurIPS 2020.

[15] David Silver, *et al. Mastering the game of go without human knowledge.* Nature 2017.

[16] Angel Chang, *et al. Shapenet: An information-rich 3d model repository.* ArXiv 2015.

[17] Agrim Gupta, *et al. LVIS: A Dataset for Large Vocabulary Instance Segmentation.* CVPR 2019.