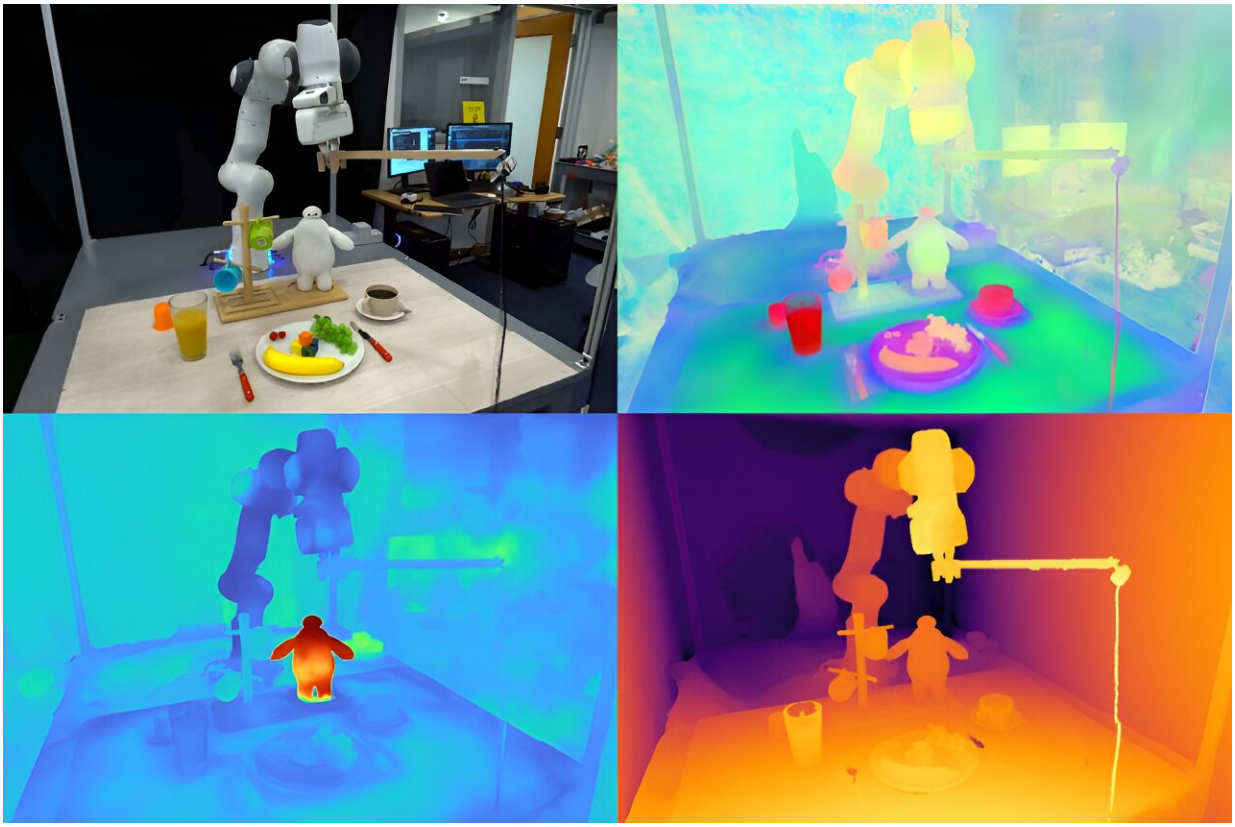
Using language to give robots a better grasp of an open-ended world

November 2 2023, by Alex Shipps



Feature Fields for Robotic Manipulation (F3RM) enables robots to interpret open-ended text prompts using natural language, helping the machines manipulate unfamiliar objects. The system's 3D feature fields could be helpful in environments that contain thousands of objects, such as warehouses. Credit: William Shen et al

Imagine you're visiting a friend abroad, and you look inside their fridge to see what would make for a great breakfast. Many of the items initially appear foreign to you, with each one encased in unfamiliar packaging and containers. Despite these visual distinctions, you begin to understand what each one is used for and pick them up as needed.

Inspired by humans' ability to handle unfamiliar objects, a group from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) designed Feature Fields for Robotic Manipulation (F3RM), a system that blends 2D images with foundation model features into 3D scenes to help robots identify and grasp nearby items. F3RM can interpret open-ended language prompts from humans, making the method helpful in real-world environments that contain thousands of objects, like warehouses and households.

F3RM offers robots the ability to interpret open-ended text prompts using natural language, helping the machines manipulate objects. As a result, the machines can understand less-specific requests from humans and still complete the desired task. For example, if a user asks the robot to "pick up a tall mug," the robot can locate and grab the item that best fits that description.

"Making robots that can actually generalize in the real world is incredibly hard," says Ge Yang, postdoc at the National Science Foundation AI Institute for Artificial Intelligence and Fundamental Interactions and MIT CSAIL. "We really want to figure out how to do that, so with this project, we try to push for an aggressive level of generalization, from just three or four objects to anything we find in MIT's Stata Center. We wanted to learn how to make robots as flexible as ourselves, since we can grasp and place objects even though we've never seen them before."

Learning 'what's where by looking'

The method could assist robots with picking items in large fulfillment centers with inevitable clutter and unpredictability. In these warehouses, robots are often given a description of the inventory that they're required to identify. The robots must match the text provided to an object, regardless of variations in packaging, so that customers' orders are shipped correctly.

For example, the fulfillment centers of major online retailers can contain millions of items, many of which a robot will have never encountered before. To operate at such a scale, robots need to understand the geometry and semantics of different items, with some being in tight spaces. With F3RM's advanced spatial and semantic perception abilities, a robot could become more effective at locating an object, placing it in a bin, and then sending it along for packaging. Ultimately, this would help factory workers ship customers' orders more efficiently.

"One thing that often surprises people with F3RM is that the same system also works on a room and building scale, and can be used to build simulation environments for robot learning and large maps," says Yang. "But before we scale up this work further, we want to first make this system work really fast. This way, we can use this type of representation for more dynamic robotic control tasks, hopefully in real-time, so that robots that handle more dynamic tasks can use it for perception."

The MIT team notes that F3RM's ability to understand different scenes could make it useful in urban and household environments. For example, the approach could help personalized robots identify and pick up specific items. The system aids robots in grasping their surroundings—both physically and perceptively.

"Visual perception was defined by David Marr as the problem of knowing 'what is where by looking,'" says senior author Phillip Isola, MIT associate professor of electrical engineering and computer science

and CSAIL principal investigator.

"Recent foundation models have gotten really good at knowing what they are looking at; they can recognize thousands of object categories and provide detailed text descriptions of images. At the same time, radiance fields have gotten really good at representing where stuff is in a scene. The combination of these two approaches can create a representation of what is where in 3D, and what our work shows is that this combination is especially useful for robotic tasks, which require manipulating objects in 3D."

Creating a 'digital twin'

F3RM begins to understand its surroundings by taking pictures on a selfie stick. The mounted camera snaps 50 images at different poses, enabling it to build a neural radiance field (NeRF), a deep learning method that takes 2D images to construct a 3D scene. This collage of RGB photos creates a "digital twin" of its surroundings in the form of a 360-degree representation of what's nearby.

In addition to a highly detailed neural radiance field, F3RM also builds a feature field to augment geometry with semantic information. The system uses CLIP, a vision foundation model trained on hundreds of millions of images to efficiently learn visual concepts. By reconstructing the 2D CLIP features for the images taken by the selfie stick, F3RM effectively lifts the 2D features into a 3D representation.

Keeping things open-ended

After receiving a few demonstrations, the robot applies what it knows about geometry and semantics to grasp objects it has never encountered before. Once a user submits a text query, the robot searches through the

space of possible grasps to identify those most likely to succeed in picking up the object requested by the user. Each potential option is scored based on its relevance to the prompt, similarity to the demonstrations the robot has been trained on, and if it causes any collisions. The highest-scored grasp is then chosen and executed.

To demonstrate the system's ability to interpret open-ended requests from humans, the researchers prompted the robot to pick up Baymax, a character from Disney's "Big Hero 6." While F3RM had never been directly trained to pick up a toy of the cartoon superhero, the robot used its spatial awareness and vision-language features from the foundation models to decide which object to grasp and how to pick it up.

F3RM also enables users to specify which object they want the robot to handle at different levels of linguistic detail. For example, if there is a metal mug and a glass mug, the user can ask the robot for the "glass mug." If the bot sees two glass mugs and one of them is filled with coffee and the other with juice, the user can ask for the "glass mug with coffee." The foundation model features embedded within the feature field enable this level of open-ended understanding.

"If I showed a person how to pick up a mug by the lip, they could easily transfer that knowledge to pick up objects with similar geometries such as bowls, measuring beakers, or even rolls of tape. For robots, achieving this level of adaptability has been quite challenging," says MIT Ph.D. student, CSAIL affiliate, and co-lead author William Shen.

"F3RM combines geometric understanding with semantics from foundation models trained on internet-scale data to enable this level of aggressive generalization from just a small number of demonstrations."

The paper, "Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation," is published on the *arXiv* preprint server.

More information: William Shen et al, Distilled Feature Fields Enable Few-Shot Language-Guided Manipulation, *arXiv* (2023). DOI: 10.48550/arxiv.2308.07931

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology