MACHINE LEARNING

SimPLE, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects

Maria Bauza[1]*, Antonia Bronars[1]*, Yifan Hou[2]†, Ian Taylor[1], Nikhil Chavan-Dafle[1], Alberto Rodriguez[1]

Existing robotic systems have a tension between generality and precision. Deployed solutions for robotic manipulation tend to fall into the paradigm of one robot solving a single task, lacking "precise generalization," or the ability to solve many tasks without compromising on precision. This paper explores solutions for precise and general pick and place. In precise pick and place, or kitting, the robot transforms an unstructured arrangement of objects into an organized arrangement, which can facilitate further manipulation. We propose SimPLE (Simulation to Pick Localize and placE) as a solution to precise pick and place. SimPLE learns to pick, regrasp, and place objects given the object's computer-aided design model and no prior experience. We developed three main components: task-aware grasping, visuotactile perception, and regrasp planning. Task-aware grasping computes affordances of grasps that are stable, observable, and favorable to placing. The visuotactile perception model relies on matching real observations against a set of simulated ones through supervised learning to estimate a distribution of likely object poses. Last, we computed a multistep pick-and-place plan by solving a shortest-path problem on a graph of hand-to-hand regrasps. On a dual-arm robot equipped with visuotactile sensing, SimPLE demonstrated pick and place of 15 diverse objects. The objects spanned a wide range of shapes, and SimPLE achieved successful placements into structured arrangements with 1-mm clearance more than 90% of the time for six objects and more than 80% of the time for 11 objects.

INTRODUCTION

Most deployed solutions for robotic manipulation fall into the paradigm of one robot, one job—like repetitively welding from point A to point B. This limits their deployment into unstructured environments like homes and hinders the automation industry from frequently and seamlessly adapting and improving manufacturing lines. For effective deployment, adaptability is essential but not sufficient. Most jobs also depend on high accuracy and reliability. We suggest then that progress in robotic manipulation deployment truly relies on achieving "precise generalization," or solving many tasks without compromising on precision. This direction brings robotic manipulation closer to the paradigm of one robot quickly adapting to and solving many jobs.

Precise pick and place enables transforming an unstructured pile of objects into a structured arrangement with objects placed in known locations, with tight tolerances. It is a challenging task because the robot needs to pick up objects that it has never interacted with before and place them accurately in target configurations. In industry, this task is commonly solved by designing specialized fixtures to localize and regrasp the part, but this can be slower and requires special-purpose fixtures.

Existing research on grasping has shown progress toward generalization by grasping many different types of objects from depth images alone (*1*–*3*). Other works have explicitly estimated object shape or pose to obtain higher-quality grasps (*4*–*6*). Although these grasping paradigms optimize grasp stability, they do not consider the grasp's usefulness to solve a downstream task (task awareness). This is a limitation for full manipulation pipelines, where grasps should be functional and stable.

Current systems that demonstrate task-aware grasping often rely on large hand-annotated datasets of task-relevant grasps (*7*–*10*), which are limited by the amount and quality of annotations. Other works relied on simulation to build task-aware metrics but were limited to narrow task definitions (*11*, *12*) or only showed that the selected grasp could be achieved without attempting the task (*13*). Fang *et al.* (*14*) focused on learning category-level task-oriented grasps for sweeping and hammering in simulation without aiming to achieve high precision. He *et al.* (*15*) used simulation for training placement-aware grasping without considering the challenges of precision placement or the uncertainty in post-grasp object pose.

In works that go beyond grasping, we find a division between achieving wide generalization and performing precise manipulations. In pick-and-place settings, approaches that consider objects with unknown geometries include learning from real experience (*16*), human annotations (*17*), or simulated data (*18*–*20*). However, these solutions tend to have wide tolerances for placements by targeting broad generality at the expense of precision.

Other works that center on precise pick and place relied on hardware adaptations or simple object shapes as well as known object geometry. For instance, the study in (*21*) achieved precise insertions by using actuator compliance, whereas the study in (*22*) combined suction and grasping to improve performance. Kleeberger *et al.* (*23*), which is most similar to our work, performed six–degree of freedom grasps to increase precision but lacked any regrasping strategy. This limits its applicability to objects and placements where there is a suitable grasp exposed.

To achieve accurate perception during manipulation, many works relied on both visual and tactile information, often referred to as visuotactile sensing. Without tactile information, postgrasp displacements are hard to estimate (*24*) and can impede precision. Zhao *et al.* (*25*) achieved precise placements through grasp selection, avoiding

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
*Corresponding author. Email: bauza@mit.edu (M.B.); bronars@mit.edu (A.B.)
†Present address: Amazon Robotics, Cambridge, MA 02142, USA.

the problem of postgrasp displacements by relying on ground-truth perception after the grasp.

With tactile sensors, robots get direct access to contacts. However, the type of contact information depends on the sensor used. Née *et al.* (*26*) found that point tactile sensors can improve grasp stability but carry limited information to estimate the object pose. Recent high-resolution sensors that rely on camera-based solutions (*27*–*30*) can facilitate accurate pose estimation (*31*–*33*) or guide manipulation of simple geometries like boxes (*34*), cables (*35*), or connectors (*36*). The solution makes use of camera-based tactile sensors and visual information to achieve accurate perception.

One existing method for leveraging tactile and visual information together for robotic manipulation is learning a joint visuotactile representation and a policy for a particular task end to end. Lee *et al.* and Chen *et al.* (*37*, *38*) fused overhead red, green, blue (RGB) images with force-torque data to learn a policy for insertion using reinforcement learning. Li *et al.* (*39*) leveraged cross-modal attention to learn a joint representation for vision (overhead RGB), touch (tactile RGB), and audio (microphone) data and then learn policies for dense packing and pouring via imitation learning. End-to-end approaches are, however, task specific and do not generalize trivially to changes in the goal specification.

Other approaches have used point clouds from vision and touch to constrain the object location for optimization-based pose estimation. Izatt *et al.* (*40*) adapted dense articulated real-time tracking (*41*) to track object pose through sequences of point clouds from vision and touch, qualitatively demonstrating that including tactile information improves tracking performance. They did not, however, evaluate the quality of pose tracking for any downstream task. Zhao *et al.* (*42*) simultaneously reconstructed the geometry of an unknown object and estimated the relative finger/object pose from a sequence of RGB and tactile images. The approach assumes that the object pose is fixed and that the initial relative finger/object pose is known.

Here, we propose to solve precise pick and place with SimPLE (Simulation to Pick Localize and placE), an approach that learns to pick, regrasp, and place objects in simulation given only the object's computer-aided design (CAD) model. It consists of three components: task-aware picking, visuotactile object pose estimation, and motion planning using hand-to-hand regrasps, as depicted in Fig. 1. We designed the components purely in simulation (Fig. 2) using the objects' geometries and then transferred SimPLE to the real system without requiring any real prior experience with the objects (Fig. 3). We demonstrated SimPLE in the context of a dual-arm robot equipped with tactile sensors and an external depth camera. From a depth image of the scene in front of the robot, we sampled grasps on the object and estimated its pose. We scored each grasp using a task-aware metric that accounts for the expected success of grasping, visuotactile object pose estimation, and pick-and-place motion planning and then executed the grasp with the highest score. Next, our visuotactile approach, which builds on our previous work on tactile perception (*31*, *32*), updated the object pose by combining the estimated pose distributions from vision (before the grasp) and tactile (after the grasp). Last, given the object pose, we executed a manipulation plan for placing the object that required solving a shortest-path problem on a graph of hand-to-hand regrasps.

In summary, this work proposes an approach to manipulation that achieves generality by only requiring known object shapes rather than expensive real robot experience with those objects and does not sacrifice precision by developing simulation tools that transfer zero shot into real setups. We experimentally demonstrated successful pick and place of 15 diverse objects (Fig. 4), and through comparisons with baseline methods and ablation studies, we validated the need for visuotactile perception and task-aware planning (Table 1). Our approach makes an important step toward enabling more flexible solutions for general-purpose robotic pick and place.

RESULTS

System overview

This work aims to solve the task of precise pick and place purely in simulation so that robots can handle a large variety of objects without requiring direct experience. We split this task into three different steps.

Task-aware grasping

From a depth image taken of the scene, SimPLE sampled antipodal grasps and computed an initial estimation of the object pose. Next, we assessed the quality of each sampled grasp using a task-aware quality metric learned in simulation and commanded the robot to execute the best antipodal grasp (Fig. 3A). Task-aware grasping refers to choosing grasps that are compatible with the task of pick and placing.

Visuotactile object pose estimation

Once an object was grasped, the robot received as input tactile observations. Combining the tactile images with the initial depth image, we updated our estimate of the distribution of possible grasped object poses (Fig. 3B).

Motion planning

Given the best estimate of the object pose, the robot computed the set of motions, including object regrasps if needed, that allowed it to place the object at a desired configuration (Fig. 3C). Last, the robot executed the motions in open loop (Fig. 3D).

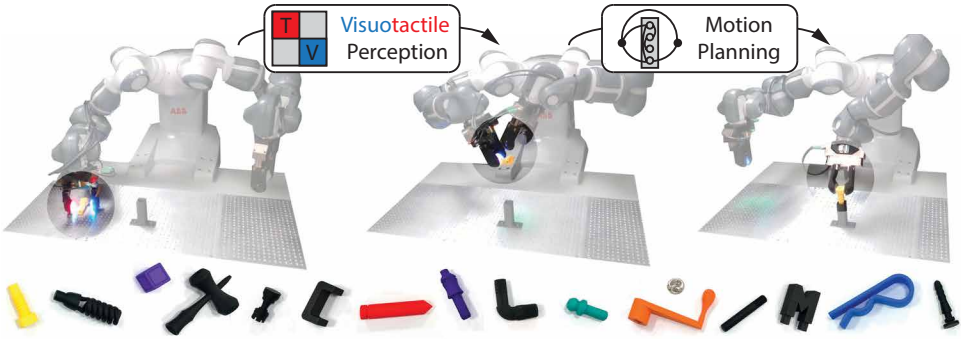


Fig. 1. Precise pick and place. We present a system capable of precisely picking and placing objects learned entirely in simulation. The proposed solution consists of three models: task-aware grasping, visuotactile perception, and motion planning. We show high-fidelity transfer of the models to the real system for the 15 objects shown at the bottom of the figure.

Learning robot models purely in simulation

To enable generality, SimPLE learns the robot models in simulation without requiring prior experience. The learned models use shape information to assess

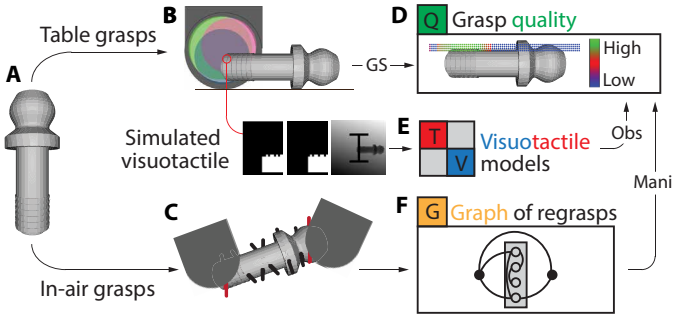


Fig. 2. Generating models in simulation. Starting from the object's CAD model (**A**), we sampled two types of grasps on the object. Table grasps (**B**) are accessible from the object's resting pose on the table. For each table grasp, we simulated corresponding depth and tactile images and used these images to learn visuotactile perception models (**E**). In-air grasps (**C**) are accessible during regrasps. We connected in-air grasp samples that are kinematically feasible into a graph of regrasps (**F**). We used the visuotactile model and regrasp graph to compute the observability (Obs) and manipulability (Mani) of a grasp and combined these with grasp stability (GS) to evaluate the quality of each table grasp (**D**).

the quality of grasps, estimate an object's pose, and compute effective motion plans for placing.

Given an object's CAD model, first, we sampled a set of grasps on the object that are accessible from the object's resting pose on the table (Fig. 2A). We denote these grasps as "table grasps" (the "Computing table-grasps" section in the Supplementary Materials provides more details on how table grasps were calculated). Next, we rendered simulated versions of the visuotactile data we expected to observe for each table grasp. This consisted of a pair of contact images (binary masks over the region of contact on each tactile sensor) and a depth image. The set of table grasps and their corresponding visuotactile data served as a library of grasps that later we could match against real observations. For each table grasp, we also evaluated and stored its task-aware quality, which corresponds to a composition of three metrics: graspability, observability, and manipulability.

The first metric, graspability, ranks different grasps depending on their capability to hold the object under disturbance forces. From the simulated contact images of each grasp, we estimated a measure in simulation of graspability where the contact region was interpreted as a contact patch with uniform pressure distribution. Intuitively, the larger the contact region, the more force and torque the grasp can resist before the contact breaks or slips.

The second metric, observability, measures the likelihood that a grasp will produce tactile observations that facilitate the estimation of the object pose within the grasp. Computing the observability of a grasp requires having a model for pose estimation. Therefore, we started by learning in simulation how to estimate an object's pose by building a visuotactile perception model tailored to that object (Fig. 2B). These models learn to match visual and tactile observations to the simulated set of visuotactile data, outputting a distribution over object poses (*31*, *32*). Details are provided in the "Visuotactile pose estimation" section in Materials and Methods.

After learning how to perceive the object, we leveraged the tactile model to evaluate the observability of each table grasp in simulation. Observability aims to quantify how informative a given contact is in uniquely determining the object pose. Because tactile sensors provide a local view of the object geometry, many contacts may look similar and therefore provide ambiguous information about the object pose. Intuitively, any grasp that looks similar to a grasp on a distant region of the object is not a good indicator of object pose and has low observability. Aiming for grasps with high observability allows a policy to prefer grasps with more unique features that, in turn, ease perception.

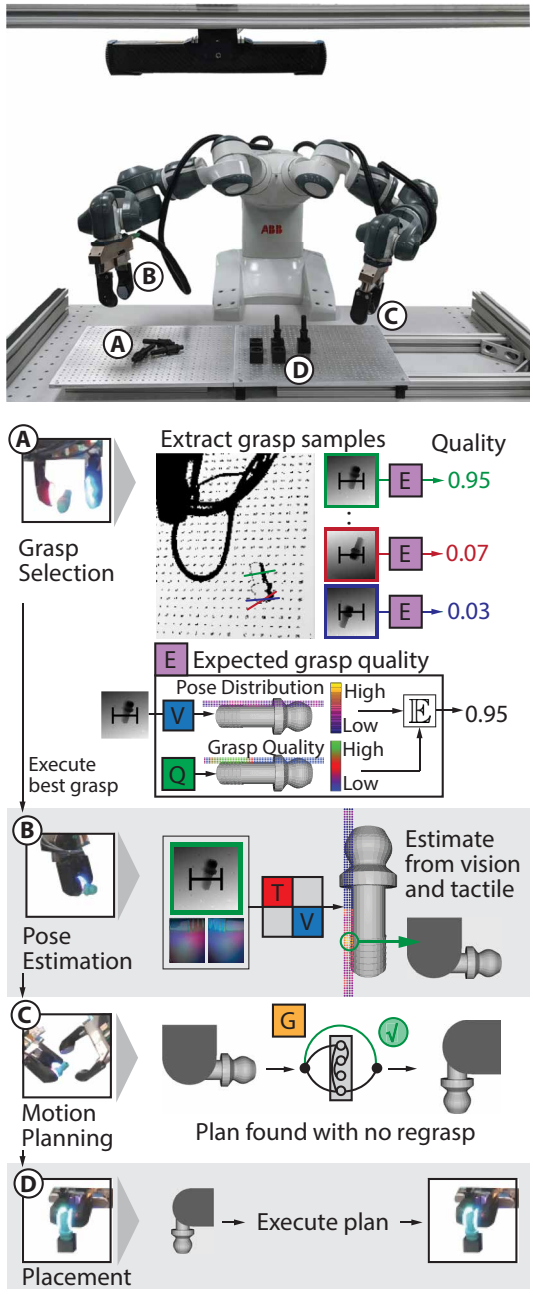


Fig. 3. Deployment in the real world. Our approach first selects the best grasp from a set of samples on a depth image (**A**). The best grasp has the highest expected quality given the pose distribution estimate from vision and the precomputed grasp quality scores. Then, we executed the best grasp and updated the pose estimate, now including information from tactile in addition to the original depth image (**B**). Next, we took the best estimate from vision and tactile as the start pose and found a plan that leads to the goal pose using the regrasp graph if necessary (**C**). Last, we executed the plan (**D**).

Object		Successes (/20 Trials)	Near-Successes (/20 Trials)	Failures (/20 Trials)
Grease		20	0	0
Head		20	0	0
Cube		19	1	0
Cotter		19	0	1
Pin		18	2	0
Pencil		18	0	2
Grip		17	3	0
Letter		17	2	1
Rod		17	1	2
Stud		16	4	0
Kendama		16	3	1
Rook		15	2	3

Fig. 4. Success rate for 15 objects. The rate of success, near-success, and failure out of 20 trials for 15 objects is tabulated in order of highest success rate to lowest.

Last, we computed the manipulability of each table grasp. Manipulability measures the simplicity of the best pick-and-place plan that we can obtain from an initial table grasp, which includes the length of motion and the number of regrasps or handing off the object to the other arm a number of times. A regrasp may be necessary to increase the workspace of the robot or if the initial grasp prevents collision-free placement of the object.

To compute this metric, we first developed a solution for finding motion plans. We started by obtaining in simulation a large set of in-air grasps (or a set of stable antipodal grasps). Then, we found possible regrasps by checking which pairs were feasible kinematically, in other words, by computing whether two grippers at each grasp in the pair would result in a collision-free situation. Precomputing the set of possible regrasps allowed us to more efficiently solve the problem of finding online a feasible placing strategy.

Finding the best motion plan for placing an object consists of building a "graph of regrasps" (*43*, *44*) (Fig. 2C) and solving for the shortest path within it. For each table grasp, we used its grasp configuration as the initial node and the desired placing pose as the goal node. Then, solving the shortest path consists of finding whether it is possible to directly place the object from its initial configuration (leading to a solution with no regrasp) or whether applying one or more regrasps will allow the robot to place the object. Edges in the regrasp graph represent kinematic feasibility between two nodes (in other words, whether a pair of grasps is feasible or whether it is possible to place the object from a given grasp without any collision). Table S1 provides the number of in-air grasps and edges of the regrasp graph of each object. Manipulability penalizes the number of required regrasps in the shortest path from a given table grasp to the goal location. Table grasps that result in plans with fewer regrasps have higher manipulability scores and are preferable.

Once we computed the three metrics in simulation, we obtained and saved the composite grasp quality for each table grasp as the product of its graspability, observability, and manipulability (Fig. 2B). In summary, simulation allowed us to compute the robot's models required to predict pose distributions using visuotactile observations; compute motion plans to place the object in a given configuration; and assess the quality of table grasps using metrics for graspability, observability, and manipulability.

Experimental evaluation of SimPLE in pick-and-place tasks

Overview of experimental setup

We validated our approach on the precise pick-and-place task, where rigid objects of different shapes need to be picked up and placed precisely on rigid fixtures (see Fig. 1). This task represents a typical application in mechanical assembly and packing. For each object tested, SimPLE first leveraged a CAD model of the object to learn purely in simulation its models for perception, grasp stability, and planning, as described in the previous section. The learned models are object specific. Our robotic system consisted of a dual arm robot with parallel jaw grippers, where tactile sensors were installed on the fingers. A top-down depth camera provided tabletop perception of the items.

Given a depth image, we followed the approach in (*2*) to find possible antipodal grasps on the image. We then filtered the grasp samples to avoid collisions between the robot's fingers, the object, and the environment. For each collision-free grasp sample, we took a crop of the original depth image that is centered and aligned at the grasp (see Fig. 3) and provided it as input to the visual model, which estimated the object pose as a distribution over table grasps.

For each sampled grasp, we computed its quality as the expectation of the quality metrics over the pose distribution provided by the visual model (Fig. 3A). The grasp sample with the highest expected quality is the most likely to provide the best combination of graspability, observability, and manipulability.

If the expected best grasp is executed successfully, the robot will also receive tactile observations. Combining the tactile observations with the previous depth image allowed us to update the vision-only estimate for the object pose (Fig. 3B). Last, we took the best visuotactile pose estimate to compute the shortest path and find the simplest motion plan that the robot can execute to place the object (Fig. 3, C and D).

We evaluated SimPLE on 15 objects (Fig. 1). For each object, we conducted 20 trials in which the robot grasped the object from an unknown starting pose and attempted to place it in a known pose.

Success metric

We categorized each trial into success, near-success, and failure. Succeeding at a pick-and-place experiment requires placing the object into a tight cavity (Figs. 1 and 3). The experiments labeled as near-success are cases where the object almost reached the goal position and failure happened because of a small misalignment (of a few millimeters) during the final placement step rather than an incorrect localization or a failed regrasp that led to complete failure.

Table 1. Success rate versus baselines on five objects. The rate of success, near-success, and failure for five objects on the proposed method and the three baselines. SimPLE and the three baselines were evaluated with 20 trials each.

Object name	SimPLE			Agnostic baseline			Tactile baseline			Vision baseline		
	S (/20)	NS (/20)	F (/20)	S (/20)	NS (/20)	F (/20)	S (/20)	NS (/20)	F (/20)	S (/20)	NS (/20)	F (/20)
Grease	20	0	0	15	0	5	19	0	1	19	0	1
Grip	17	3	0	10	2	8	1	0	19	18	2	0
Rod	17	1	2	3	8	9	12	5	3	15	0	5
Kendama	16	3	1	15	2	3	10	3	7	14	6	0
Hose	14	3	3	15	3	2	8	2	10	12	6	2

Near-successes could, in principle, be resolved by a closed-loop local insertion strategy (see "Analysis of the near success" section in the Supplementary Materials for an in-depth analysis of near successes). Any other outcome was considered a failure.

Baseline experiments

For 5 of the 15 objects, we also conducted a set of baseline experiments to evaluate the influence of each of the core components of SimPLE: task-aware grasping, tactile localization, and visual localization. Each baseline eliminated one of these components but preserved the other two. In the tactile baseline, after performing a task-aware grasp, pose estimation only used tactile information instead of using both visual and tactile observations. In the vision baseline, after performing a task-aware grasp, pose estimation only used vision information instead of using both visual and tactile observations. Last, for the task-agnostic baseline, instead of using the task-aware grasp selection, this baseline selected grasps on the basis of the grasp quality metric from Dex-Net (*2*). Perception for this baseline used both visual and tactile information to estimate the object pose after the grasp.

SimPLE versus tactile baseline

Tactile localization performs well in combination with task-aware grasping ("tactile baseline") for objects where unambiguous grasp locations exist. The success rate for "grease," for example, was 95% of 20 trials, and for "rod," the tactile baseline achieved 60% successes and 25% near-successes out of 20 trials (Table 1). Some objects, such as "hose," "grip," and "kendama," did not have any region where the grasp was entirely unambiguous. Therefore, even in the presence of task-aware grasping, tactile localization alone is unable to consistently resolve the object pose. This was reflected by a lower rate of successes and near-successes for the tactile baseline for these objects (Table 1).

SimPLE versus vision baseline

Visual localization performs well in combination with task-aware grasping ("vision baseline") when the object configuration is unambiguous from an overhead camera. This was the case for hose, grip, kendama, and grease, whose rates of successes and near-successes were correspondingly high (Table 1). The object rod, on the other hand, had one threaded end and one unthreaded end, which were difficult to disambiguate using only an overhead camera. This caused the vision baseline to fail in 25% of the 20 trials. The rate of near-successes for hose and kendama was high relative to the rate of true successes. This is because although vision is capable of globally disambiguating between object orientations, it is less adept at refining the object pose to high precision, which is necessary to successfully place the object into the cavity. When tactile localization was included, we saw more true successes (the "Comparison of tactile pose estimation versus tactile pose estimation with a prior" section in the Supplementary Materials provides additional analysis on the influence of tactile versus vision on pose estimation accuracy).

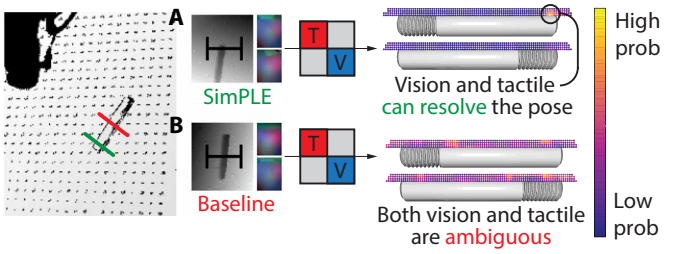
SimPLE versus task-agnostic baseline

Next, we compared SimPLE against a "task-agnostic baseline," which uses the quality metric from Dex-Net (*2*) and therefore is task agnostic and object independent. Note that SimPLE takes into account the downstream goals of localization and placement, in addition to object-tailored graspability, before choosing a grasp. Instead, the task-agnostic baseline considers only the likelihood of grasp success when choosing a grasp. For some objects, like hose and kendama, the task-agnostic baseline often prefers the same grasp as SimPLE, and therefore, the success rates are similar. For other objects, like rod and grease, task awareness leads SimPLE to prefer a different grasp than the task-agnostic baseline.

For the object rod, we found that task-aware grasping facilitates perception by targeting grasps that lead to unambiguous tactile information (Fig. 5). This is particularly important for rod because visual information alone is ambiguous. The SimPLE method resulted in 85% successful trials out of 20 compared with 15% successful trials out of 20 for the task-agnostic baseline. The task-agnostic baseline targets grasped near the center of the object, where both vision and tactile information are ambiguous. SimPLE was able to resolve the pose by targeting grasps near the end of the object, where tactile information can resolve the ambiguity in the visual information.

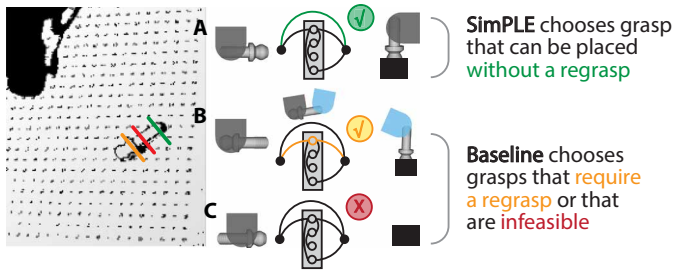
The object rod also highlights an important benefit of our regrasp-based planning framework: It allows for recovery from vision-based perception failures. In 6 of SimPLE's 17 successful trials, the robot grasped on the wrong end of rod because of a vision ambiguity (the two ends of the object are difficult to disambiguate from vision alone). After incorporating tactile information, the robot amended its estimate and planned a regrasp to place the object in the correct orientation. The SimPLE-preferred grasp for this object was expected to result in a placement without a regrasp. However, the vision confusion and the consequent correction from tactile information triggered a correction regrasp by the motion planning framework that allowed the robot to recover from this vision failure.

As another example, for the object grease, we found that task-aware grasping facilitated motion planning by targeting grasps that led to successful motion plans (Fig. 6). Because grease is a small object, regrasps were infeasible from some table grasps. This can make it impossible to place the object if motion planning is not considered



	Successful trials
SimPLE (task-aware)	17/20
Baseline (task-agnostic)	3/20

Fig. 5. Task-aware grasping facilitates perception. We consider the case of the object rod and compare SimPLE against a baseline that does not perform task-aware grasping. We show the type of grasp selected by SimPLE (**A**) and the baseline (**B**) and the final pose distribution after grasping the object. The baseline could not resolve the pose because both tactile and vision observations were ambiguous. Instead, SimPLE aimed at grasps that it expected would produce observable tactile imprints, allowing the perception model to resolve the pose after the grasp. As a result, SimPLE succeeded in 85% of 20 real experiment trials, whereas the baseline, because of perception errors, only succeeded in 15% of 20 trials.



	Successful trials	Number regrasps
SimPLE (task-aware)	20/20	0: 20
Baseline (task-agnostic)	15/20	0: 1, 1: 14

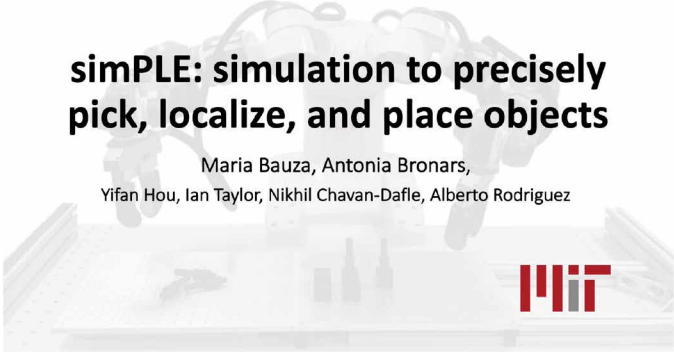
Fig. 6. Task-aware grasping results in better planning solutions. A baseline that does not aim at task-aware grasping can end up choosing grasps that require more regrasps than needed (**B**) or that result in an object configuration where no motion plan is possible (**C**). Instead, our method chose grasps that reduced the number of regrasps required to place the object (**A**). Aiming for grasps that require simpler planning strategies ended in more successful trials.

before the first grasp. SimPLE resulted in 100% successful trials out of 20 compared with 75% successful trials out of 20 for the task-agnostic baseline. The task-agnostic baseline aimed at grasps near the center of the object that require a regrasp. In the 25% of failed trials, the initial grasp did not leave enough room for a regrasp, and thus the attempt failed. The remaining 75% of trials of the task-agnostic baseline were successful, but they required a regrasp that SimPLE avoided. SimPLE targeted grasps that require simpler planning strategies and, as a consequence, resulted in a higher success rate.

Summary of SimPLE's performance

We tested SimPLE on a total of 15 objects to evaluate the method's generalization to a wide variety of object shapes and sizes. Figure 4 shows the number of attempts that succeeded after 20 trials, as well as the number of near-successes and failures. Movie 1 shows a successful placement for each object as well as a subset of failures and near-successes. Overall, SimPLE provides the robot with a method that successfully transfers to the real world and achieves precise pick and place. For nine of the objects, SimPLE had a success rate of at least 85%. Only "magnet" had a success rate of less than 50%. The object magnet is particularly challenging because there are very few object orientations and grasps that make its pose observable.

Of the several objects studied, we highlight the case of "stud" because it provides an important discussion point for SimPLE. As depicted in Fig. 7, the best table grasps for stud, according to the task-aware quality, required at least one regrasp to ensure tactile observability. During experiments, SimPLE selected these grasps and had a success rate of 80%, with 20% near-successes and no failures. Therefore, all selected table grasps required a regrasp. Choosing other table grasps would not require a regrasp but would come at the cost of losing tactile observability and thus a higher likelihood of failing to estimate an accurate object pose. Dealing with such trade-offs is inevitable and requires foresight, because task awareness implies balancing multiple objectives (graspability, observability, and manipulability).



Movie 1. Overview of the SimPLE approach and results. The video highlights the main advantages of SimPLE, shows the method step by step, and demonstrates a successful placement for each object. It also shows examples of consecutive placements and representative failure cases.

DISCUSSION

In this work, we present an approach to precise pick and place that leverages offline computing to allow a high degree of adaptability. Being able to use the same algorithms and system to precisely place objects taken from unstructured scenes is at the core of many manipulation problems that still remain unsolved. By not requiring prior robot or human experience with those objects, we showcase that it is possible to aim to build systems that rapidly adapt to additional objects without compromising on accuracy.

Our experiments suggest that through visuotactile perception and task awareness, robot models learned in simulation can successfully transfer to real systems. We showed this for a large variety of objects' sizes and shapes. When compared with baselines, our results validate the need for having both tactile and visual sensing, as well as the benefits of deploying policies that consider the task requirements end to end.

Simulated observations to learn robot models

Although it is common to learn behavior policies purely in simulation, here, we used simulation to learn precise perception models that

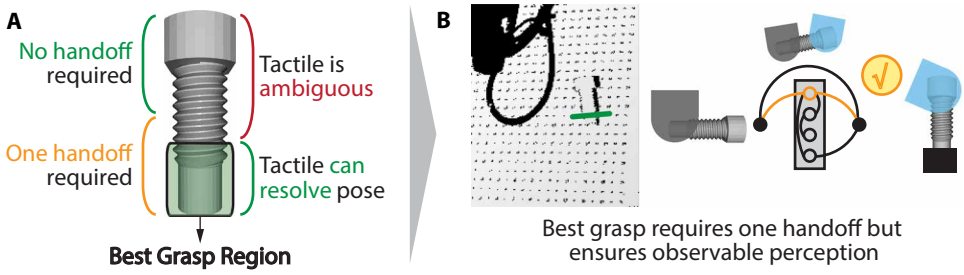


Fig. 7. Task-aware grasping balances perception and planning to find the grasp region that maximizes success. For the object stud, there was no one grasp region that maximized both observability and manipulability (**A**). The proposed approach chose a grasp that require a handoff to ensure observable perception, negotiating the trade-off between observability and manipulability to maximize task success (**B**).

can then inform model-based behavior policies (the "Tactile localization with real versus simulated contact shapes" section in the Supplementary Materials provides additional analysis on the pose estimation accuracy on real versus simulated data). Learning perception purely in simulation allowed us to generate the large amounts of data needed to predict meaningful pose distributions. By simulating observations, we were also able to rethink perception as a matching problem that aims to answer how similar observations are rather than directly regress which pose could produce a given observation. Matching is simpler and allowed us to easily generate probabilities that quantified the similarity of different observations on the basis the proximity of the contacts that generated them. Combining vision, which is global, with tactile, which is local but provides higher resolution at contact, makes our system capable of precisely estimating poses for objects with more variety in shapes, textures, and sizes.

Task awareness to select grasps

Our experiments suggest that accounting only for the success of a grasp is less effective than selecting task-aware grasps that also account for facilitating perception and planning. We show this in more detail in Figs. 5 and 6, where selecting task-aware grasps was key to resolving the object pose and to finding placing plans with fewer regrasps, respectively. During experiments, we measured the task-aware quality of a grasp by taking the product of three metrics: graspability, observability, and manipulability. Nevertheless, SimPLE would allow us to seamlessly integrate other relevant metrics like penalizing grasps on deformable or fragile regions of an object. To further improve our solution, we believe that it would be useful to systematically assess the importance of each metric toward solving the task. This could lead to better ways of defining and combining metrics. For instance, ensuring good pose estimation might be of higher or lower importance than simplifying a motion plan for end-to-end reliability.

Scope of applicability

SimPLE can accommodate different sensor modalities as long as it is possible to simulate the sensor observations. Following the process that SimPLE uses for our tactile and vision sensors, we would learn a perception model on the basis of matching real observations to a set of precomputed ones. Because each model outputs a distribution of likelihoods over possible poses, integrating the sensing modalities reduces to simply multiplying the likelihoods obtained from each of the observations (*31*, *32*).

Also relevant to the applicability of SimPLE, this method can integrate task constraints such as avoiding environment collisions or dealing with multiple objects on the scene. Qualitative experiments have shown that SimPLE for grasp selection is able to find good grasps on the objects even when multiple objects (of different types) are present. In the cases where there are multiple types of objects, the robot can run grasp selection for each object type and select the grasp with highest quality that solves the object classification problem implicitly.

Opportunities for future work

SimPLE shows that it is possible to build adaptable solutions for robotic manipulation while achieving high accuracy. To that end, SimPLE features an open-loop solution based on learning from simulated data. Next, we expose three of its limitations and comment on how to overcome them.

Near-success failures

As described in Results and the "Analysis of the near success" section in the Supplementary Materials, it was not uncommon to have near success because of small misalignments that ended in the object being placed quite close yet outside the ±0.5-mm placing tolerance. Such cases would benefit from closed-loop execution during the placement. For instance, we could measure the forces experienced during placing and detect external contacts (*45*). This could allow the robot to identify the evolution of the placement and derive a policy for correcting deviations.

Reacting to the unexpected

SimPLE is currently executed in open loop, meaning that after receiving tactile images from the initial grasp and producing a visuotactile pose estimate, it does not update its belief of the object pose. Although we showed that this is sufficient to get relatively good results, in some cases, it proves insufficient. Adding a tracking strategy that aggregates multiple sensor readings over time would help avoid cases of ambiguity and reduce failure. A step in the system that would benefit from additional feedback is after regrasps, which tend to exacerbate error. Such a strategy could also help deal with unexpected disturbances like the object sliding in the grasp because of collisions with a dynamic or unknown environment.

Defining observability under multiple sensors

Currently, the notion of observability only takes into account how likely a grasp is to produce useful observations for tactile perception. Although we took both vision and tactile into account when estimating the object pose, to compute observability, we currently do not account for vision, which in some cases would be sufficient to estimate the pose even if tactile would be ambiguous. We believe that building the algorithmic and mathematical tools to better understand and exploit observability is paramount. Defining observability after aggregating multiple sensors would increase the range of poses where the object pose can be estimated. Moreover, it would also be possible to provide the robot with policies that take multiple actions, rather than a single grasp as in SimPLE, to ensure higher observability and thus avoid wide or multimodal pose distributions.

SimPLE leverages object-specific perception models learned entirely in simulation to perform precise pick and place without any real experience with the objects. In this way, SimPLE is an approach to manipulation that achieves generality by only requiring known object shapes rather than expensive robot experience. We experimentally validated SimPLE's capability via successful pick and place of 15 objects of diverse shapes and sizes, resulting in successful placements more than 90% of the time for six objects and more than 80% of the time for 11 objects. Our approach therefore makes an important step toward enabling more flexible solutions for general-purpose robotic pick and place.

MATERIALS AND METHODS

Notation for table and in-air grasps

We considered as a valid table grasp, t, a grasp that could occur without collisions between the robot's fingers, a flat environment like a table, and an object resting on top of it. We denoted as T a fixed discrete set taken from all the possible table grasps. Given a table grasp t_o that produced the observation o, we were interested in computing the density function

$$p(t = \text{argmin}_{t' \in T}\ dist(t', t_o) \in T | o) \quad (1)$$

which represents how likely a grasp $t \in T$ is to be the closest in pose distance (measured by the function *dist*) to a table grasp t_o. To simplify notation, we will denote the probability above as $p(t|o)$. The *dist* function corresponds to the ADD (average three-dimensional distance) (*46*) metric, which measures the average distance between the point cloud of the object in two different poses.

In our case, an observation o of a table grasp t included a simulated visual observation from the depth camera in the form of a depth image, d, and a pair of simulated tactile images from the tactile sensors, c, such that $o = (d, c)$. Our perception models then estimate the following distribution:

$$p(t|o) = p(t|d, c) \quad (2)$$

which provides the likelihood of each table grasp in T given the visuotactile observations d and c (which are simulated observations during learning and real observations at test time).

For each table grasp, we developed functions to simulate several of its attributes such as the expected contacts $\hat{c}(t)$ on each finger of the grasp, simulated depth image $\hat{d}(t)$ of the grasp (see Fig. 2), and quality score, $q(t)$. Each table grasp t is uniquely defined by considering the pose of the object with respect to the gripper frame. We will also refer to this pose as t to simplify notation.

We also considered in-air grasps, denoted as g, which correspond to grasps that could happen without collisions between the robot fingers and the object. Note that then the environment is not constraining the grasp, and thus in-air grasps include but are not restricted to being table grasps. For in-air grasps, we considered the notion of regrasps, r, to denote whether two grippers grasping an object at the grasp locations g_1, g_2 would create any collisions. Therefore, $r(g_1, g_2)$ is a binary variable that is equal to 0 if there is some collision during the regrasp and 1 if the regrasp is collision free. With this definition of regrasp, we considered a discretized set of in-air grasps, G, and built a graph of regrasps, $\mathbf{G} = \mathbf{G}(t_{\text{init}}, q_{\text{goal}})$, where its nodes correspond to all the in-air grasps g from G, as well as the initial table grasp, t_{init}, and the desired goal configuration of the object at placement, q_{goal}. The edges between any grasps $g_1, g_2 \in \mathbf{G}$ on the graph $\mathbf{G}$ are precomputed and exist if $r(g_1, g_2) = 1$. At runtime, start (t_{init}) and goal (q_{goal}) nodes are added, as well as edges that connect them to the precomputed graph.

Table grasps

Table grasps are the set of grasps that could occur without collision between the robot fingers, a flat environment like a table, and an object resting on top of that table. We generated a discrete set T of table grasps by first determining the resting configurations of the object on a flat table. Then, we sampled antipodal grasps with 1-mm resolution around the object. We constrained the height of the antipodal samples by enforcing that the tips of the robot fingers were flush with the table during the grasp. We augmented the set of table grasps by perturbing the object from its resting configuration to account for any object perturbations that may occur during grasping. We considered small perturbations about the axis of the grasp and of the height of the grasp. The set of table grasps T consists of poses, t, measured relative to the gripper and the corresponding set of simulated visuotactile data for those poses, $\hat{c}(t)$ and $\hat{d}(t)$. The process for rendering visuotactile data in simulation is described below. The "Computing table-grasps" section of the Supplementary Materials provides more details on how table grasps are computed.

Rendering observations in simulation

We used a similar procedure to simulate contact and depth images to represent the actual observations from real sensors. How to simulate contact is extensively described in (*31*, *32*) and consists of rendering images from a virtual camera of an object placed such that its closest point to the virtual camera would contact without penetrating an imaginary sensor. The virtual camera rendered a depth image from the scene, which we then processed to create a contact mask using the pixels where the object would penetrate the sensor.

Simulating depth images requires rendering both the table and object at a given position and taking a depth image using a virtual camera that matches the extrinsic and intrinsic parameters of the real one. We computed the depth image associated with each table grasp by taking a crop of the rendered depth image by centering and reorienting it at the pose of the table grasp (*32*). The resulting crop and the contact images simulated for each of the fingers on the gripper provided the set of simulated observations for each table grasp.

Visuotactile pose estimation

For each object and sensor modality (vision and tactile in our case), we trained an encoder to match observations to the precomputed set of observations in the set of table grasps T, following the approach outlined in (*32*). Once we learned to encode observations on the basis of their distances in pose space, we precomputed encodings for the observations in the set T. At test time, we computed an encoding for the real observation coming from the actual sensor and compared its distance in embedding space with all of the precomputed encodings in T. We applied a softmax to the resulting vector of distances to obtain a probability distribution over the table gasps in T. This distribution represents the likelihood of the object configuration matching each element in T given the real observation.

For each table grasp t, we precomputed encodings for its two contact images and the depth image that represents the expected observation at

that grasp. This allowed us to efficiently compute the distributions over possible object poses for each sensing modality online.

During experiments, visuotactile observations consisted of a depth image, d, of the object before grasping, two contact images c, and the gripper width during a grasp. The final estimate of the pose distribution, $p(t|d, c)$, came from the product of the distribution obtained from the depth image $p(t|d)$, the distributions obtained from the contact images $p(t|c)$, and a Gaussian centered at the gripper width. Additional information on visuotactile pose estimation, including training details, is available in the "Visuotactile pose estimation" section in the Supplementary Materials.

In-air grasps and finding object regrasps

In-air grasps, g, were selected by dividing the object model surface into small patches, computing the normals of each patch, and pairing the patches to compute if they could result, within some tolerance, in a stable antipodal grasp (*44*, *47*).

We also computed whether a regrasp $r(g_1, g_2)$ was possible between grasps g_1, g_2 by checking whether a gripper taking grasp g_1 and another at g_2 would collide between them. If not, $r(g_1, g_2) = 1$ and 0 otherwise. Precomputing the set of regrasps for all pairs in G made motion planning more efficient when building $\mathbf{G}$.

Regrasp graph and shortest-path search

To construct the regrasp graph $\mathbf{G}$, we used the set of grasps G, the initial table grasp t_{init}, and the goal configuration q_{goal}. Although edges between grasps in G were precomputed in simulation, the rest of the edges were computed online. To simplify computations, we ran a shortest-path search on the graph to find the simplest plan to the goal, which represents the minimal number of regrasps. First, we checked whether t_{init} could happen without collisions with the robot and environment when the object was at q_{goal}. If so, o regrasp was needed to place the object.

Otherwise, we computed for each grasp $g \in G$ if $r(t_{\text{init}}, g) = 1$. If so, we added that edge to the graph. We did the same for q_{goal}, checking whether it was possible to exert a grasp $g \in G$ if the object was at q_{goal}. Next, we solved a shortest-path problem where each edge was modulated by the regrasp quality, thus preferring plans that were more likely to succeed. The "Building a regrasp graph" section in the Supplementary Materials provides more details on how the regrasp graph was built.

Task-aware grasp quality

For each table grasp $t \in T$, we computed in simulation its grasp quality, $q(t)$, as a task-aware metric that consists of the product of three measures.

Graspability

Graspability measures the stability of a grasp to hold within the robot's fingers. We computed graspability for each table grasp t by normalizing the area of the contact patches from the simulated observations c_t to fall between 0 and 1. A graspability of 1 indicates that a grasp was expected to be stable.

Observability

Estimating the pose of an object tactilely can be unambiguous, making it preferable to grasp at regions with more unique features. We quantified this intuition with the metric of tactile observability. To obtain the observability of a grasp t, we computed how likely each table grasp $t' \in T$ was given the simulated observations from t, $p(t'|d_t, c_t)$. Then, we checked whether the most likely table grasp was within 5 mm of t and whether the deviation between the five most likely table grasps was less than 2 mm. If so, t was observable, and we denoted its observability as 1 and 0 otherwise.

Manipulability

For each table grasp t, we solved the shortest path for $\mathbf{G}(t, q_{\text{goal}})$, which resulted in the simplest plan to place the object. Manipulability score was 1 if no regrasp was needed, 0.8 or 0.4 if one or two regrasps were needed, respectively, and 0 otherwise.

To ensure consistency between adjacent table grasps, we smoothed the scores by computing for a given table grasp its closest table grasps and updated its scores as the minimum of its original score and the median and mean of its closest scores. Two grasps were close if their angle distances were smaller than 5° and less than 10 mm in the x and y directions. We empirically determined the numerical values used to assess grasp quality, such as the 2-mm and 5° thresholds.

Pipeline for real experiments

To deploy SimPLE during real experiments, we first sampled possible antipodal grasps, x, without knowledge of the object position. Then, we scored each sample by computing its expected score as

$$E[q(x)|d] = \Sigma_t \, q(t) \cdot p(t|d) \tag{3}$$

The sample with the highest expected quality was selected and executed. This in turn allowed us to get tactile images c and update our estimate over the table grasp exerted: $p(t|d, c)$.

To place the object, we took $t_{\text{init}} = \arg\max_t p(t|d, c)$ and computed the graph $\mathbf{G}(t_{\text{init}}, q_{\text{goal}})$ by checking the existence of the edges between t_{init} and G, t_{init} and q_{goal}, and between G and q_{goal}. Last, we found the shortest path in $\mathbf{G}$, which provides the motion plan that the robot will execute to place the object.

Grasp sampling

Our grasp sampling strategy was based on (*2*), which uses a depth image from the scene to identify possible antipodal grasps. We extended it by also checking whether the fingers were likely to collide with the environment. All baselines also made use of this sampling strategy.

Robot system setup

The robot system we used to conduct real experiments consisted of a dual-arm ABB Yumi with two WSG-32 grippers and GelSlim 3.0 tactile (*30*) sensing fingers to collect tactile observations. We used a Photoneo PhoXi M depth camera mounted overhead to collect visual observations. We placed objects into cavities that were a negative of the object with 1 mm of radial clearance from the object and a 3-mm chamfer on the mating edge.

Success, near-success, and failure during experiments

In successful trials, the robot placed the object into the cavity. In nearly successful trials, the object was in the correct location and orientation but marginally missed the cavity. Intuitively, these attempts could have succeeded with a local insertion strategy. Failed trials included grasp failures, global localization failures where the object pose estimate was flipped in the wrong orientation, and motion planning failures where a handoff was unsuccessful or no feasible plan was found from the initial grasp. Failed trials could not have been salvaged by a local insertion strategy.

Baseline description

We compared the performance of our system against three baselines.

Agnostic baseline

We scored the set of grasp candidates on the basis of a task-agnostic and object-independent Grasp Quality Convolutional Neural Networks (GQ-CNN) (*2*). This quality network was trained to predict the robustness of a candidate grasp without considering the downstream task or the object grasped. To evaluate the grasp, the agnostic baseline took as input the gripper depth and the same image that SimPLE uses, consisting of a depth image centered on the grasp center pixel and aligned with the grasp axis orientation. Its output was a robustness score for the candidate grasp between 0 and 1. We executed the grasp with the highest score. The visuotactile localization step and the motion planning were the same as in SimPLE.

Tactile baseline

The grasp selection step was the same as in SimPLE, but the visuotactile localization step was replaced with tactile localization alone, $p(t \mid c)$. After executing the best grasp, we found the pose that maximizes the joint distribution from both contacts and a Gaussian centered at the measured gripper width. We took that maximizing pose as the tactile pose estimate to compute the motion plan.

Vision baseline

The grasp selection step was the same as in SimPLE, but the visuotactile localization step was replaced with vision localization alone, $p(t|d)$. The pose estimate used for motion planning was then the pose that maximized the distribution from vision.

Supplementary Materials

This PDF file includes:

Figs. S1 and S2
Tables S1 to S3
References (*48*–*50*)

Other Supplementary Material for this manuscript includes the following:

Movie S1
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. A. Zeng, S. Song, K. T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. Chavan-Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, A. Rodriguez, Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *Int. J. Rob. Res.* **41**, 690–705 (2022).
2. J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, K. Goldberg, Learning ambidextrous robot grasping policies. *Sci. Robot.* **4**, eaau4984 (2019).
3. S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, K. Bousmalis, Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 12627–12637.
4. D. Chen, V. Dietrich, Z. Liu, G. Von Wichert, A probabilistic framework for uncertainty-aware high-accuracy precision grasping of unknown objects. *J. Intell. Robot. Syst.* **90**, 19–43 (2018).
5. X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, D. Fox, Self-supervised 6d object pose estimation for robot manipulation, in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2020), pp. 3665–3671.
6. N. Chavan-Dafle, S. Popovych, S. Agrawal, D. D. Lee, V. Isler, Simultaneous object reconstruction and grasp prediction using a camera-centric object shell representation, in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2022), pp. 1396–1403.
7. C. Yang, X. Lan, H. Zhang, N. Zheng, Task-oriented grasping in object stacking scenes with crf-based semantic model, in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2019), pp. 6427–6434.
8. L. Manuelli, W. Gao, P. Florence, R. Tedrake, kpam: Keypoint affordances for category-level robotic manipulation, in *Robotics Research: The 19th International Symposium ISRR* (Springer International Publishing, 2022), pp. 132–157.
9. A. Murali, W. Liu, K. Marino, S. Chernova, A. Gupta, Same object, different grasps: Data and semantic knowledge for task-oriented grasping, in *Conference on Robot Learning* (MLResearchPress, 2021), pp. 1540–1557.
10. M. Sun, Y. Gao, GATER: Learning grasp-action-target embeddings and relations for task-specific grasping. *IEEE Robot. Autom. Lett.* **7**, 618–625 (2022).
11. X. Lou, Y. Yang, C. Choi, Collision-aware target-driven object grasping in constrained environments, in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 6364–6370.
12. K. Xu, H. Yu, R. Huang, D. Guo, Y. Wang, R. Xiong, Efficient object manipulation to an arbitrary goal pose: Learning-based anytime prioritized planning, in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 7277–7283.
13. B. Wen, W. Lian, K. Bekris, S. Schaal, Catgrasp: Learning category-level task-relevant grasping in clutter from simulation, in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 6401–6408.
14. K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, S. Savarese, Learning task-oriented grasping for tool manipulation from simulated self-supervision. *Int. J. Robot. Res.* **39**, 202–216 (2020).
15. Z. He, N. Chavan-Dafle, J. Huh, S. Song, V. Isler, Pick2Place: Task-aware 6DoF grasp estimation via object-centric perspective affordance. arXiv:2304.04100 [cs.RO] (8 April 2023).
16. L. Berscheid, P. Meißner, T. Kröger, Self-supervised learning for precise pick-and-place without object model. *IEEE Robot. Autom. Lett.* **5**, 4828–4835 (2020).
17. H. Chen, T. Kiyokawa, W. Wan, K. Harada, Category-association based similarity matching for novel object pick-and-place task. *IEEE Robot. Autom. Lett.* **7**, 2961–2968 (2022).
18. M. Gualtieri, R. Platt, Learning 6-dof grasping and pick-place using attention focus, in *Conference on Robot Learning* (MLResearchPress, 2018), pp. 477–486.
19. M. Gualtieri, A. T. Pas, R. Platt, Pick and place without geometric object models, in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 7433–7440.
20. M. Gualtieri, R. Platt, Robotic pick-and-place with uncertain object instance segmentation and shape completion. *IEEE Robot. Autom. Lett.* **6**, 1753–1760 (2021).
21. A. S. Morgan, B. Wen, J. Liang, A. Boularias, A. M. Dollar, K. Bekris, Vision-driven compliant manipulation for reliable, high-precision assembly tasks. arXiv:2106.14070 [cs.RO] (26 June 2021).
22. K. Kleeberger, J. Schnitzler, M. U. Khalid, R. Bormann, W. Kraus, M. F. Huber, Precise object placement with pose distance estimations for different objects and grippers, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021), pp. 4639–4646.
23. K. Kleeberger, M. Völk, M. Moosmann, E. Thiessenhusen, F. Roth, R. Bormann, M. F. Huber, Transferring experience from simulation to the real world for precise pick-and-place tasks in highly cluttered scenes, in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2020), pp. 9681–9688.
24. J. Zhao, J. Liang, O. Kroemer, Toward precise robotic grasping by probabilistic post-grasp displacement estimation, in *Field and Service Robotics: Results of the 12th International Conference* (Springer, 2021), pp. 131–144.
25. J. Zhao, D. Troniak, O. Kroemer, Towards robotic assembly by predicting robust, precise and task-oriented grasps, in *Conference on Robot Learning* (MLResearchPress, 2021), pp. 1184–1194.
26. N. Nikandrova, E. Kolycheva, V. Kyrki, Task-specific grasping of similar objects by probabilistic fusion of vision and tactile measurements, in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)* (IEEE, 2015), pp. 704–710.
27. W. Yuan, S. Dong, E. H. Adelson, Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors* **17**, 2672 (2017).
28. N. F. Lepora, Soft biomimetic optical tactile sensing with the TacTip: A review. *IEEE Sens. J.* **21**, 21131–21143 (2021).
29. M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, R. Calandra, Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robot. Autom. Lett.* **5**, 3838–3845 (2020).
30. I. H. Taylor, S. Dong, A. Rodriguez, Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger, in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 10781–10787.
31. M. Bauza, A. Rodriguez, B. Lim, E. Valls, T. Sechopoulos, Tactile object pose estimation from the first touch with geometric contact rendering, in *Conference on Robot Learning* (MLResearchPress, 2021), pp. 1015–1029.
32. M. Bauza, A. Bronars, A. Rodriguez, Tac2pose: Tactile object pose estimation from the first touch. *Int. J. Rob. Res.* **42**, 1185–1209 (2023).

33. P. Sodhi, M. Kaess, M. Mukadam, S. Anderson, Learning tactile models for factor graph-based estimation, in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 13686–13692.
34. F. R. Hogan, J. Ballester, S. Dong, A. Rodriguez, Tactile dexterity: Manipulation primitives with tactile feedback, in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2020), pp. 8863–8869.
35. Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, E. Adelson, Cable manipulation with a tactile-reactive gripper. *Int. J. Robot. Res.* **40**, 1385–1401 (2021).
36. R. Okumura, N. Nishio, T. Taniguchi, Tactile-sensitive NewtonianVAE for high-accuracy industrial connector-socket insertion. arXiv:2203.05955 [cs.RO] (10 March 2022).
37. M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, J. Bohg, Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Trans. Robot.* **36**, 582–596 (2020).
38. Y. Chen, M. Van der Merwe, A. Sipos, N. Fazeli, Visuo-tactile transformers for manipulation, in *Conference on Robot Learning* (MLResearchPress, 2023), pp. 2026–2040.
39. H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, J. Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation, in *Conference on Robot Learning* (MLResearchPress, 2023), pp. 1368–1378.
40. G. Izatt, G. Mirano, E. Adelson, R. Tedrake, Tracking objects with point clouds from vision and touch, in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), pp. 4000–4007.
41. T. Schmidt, R. A. Newcombe, D. Fox, DART: Dense articulated real-time tracking, in *Robotics: Science and Systems 2014* (RSS, 2014), pp. 1–9.
42. J. Zhao, M. Bauza, E. H. Adelson, FingerSLAM: Closed-loop unknown object localization and reconstruction from visuo-tactile feedback, in *2023 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 8033–8039.
43. W. Wan, M. T. Mason, R. Fukui, Y. Kuniyoshi, Improving regrasp algorithms to analyze the utility of work surfaces in a workcell, in *2015 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2015), pp. 4326–4333.
44. Y. Hou, J. Zhenzhong, M. T. Mason, Fast planning for 3d any-pose-reorienting using pivoting, in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 1631–1638.
45. S. Kim, A. Rodriguez, Active extrinsic contact sensing: Application to general peg-in-hole insertion, in *International Conference on Robotics and Automation* (ICRA) (IEEE, 2022), pp. 10241–10247.
46. B. Tekin, S. N. Sinha, P. Fua, Real-time seamless single shot 6d object pose prediction, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 292–301.
47. I.-M. Chen, J. W. Burdick, Finding antipodal point grasps on irregularly shaped objects. *IEEE Trans. Robot. Autom.* **9**, 507–512 (1993).
48. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
49. K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 9729–9738.
50. P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 1125–1134.

Acknowledgments: We thank F. Alet for the insightful discussions related to the project. **Funding:** The research was supported by ABB and Magna International. **Author contributions:** M.B. devised the main solution of the paper, integration of the full system, designed and executed experiments, and wrote the manuscript. A.B. took on the perception system, executed experiments, and helped write the manuscript. Y.H. designed and implemented the motion and regrasp planning foundation. N.C.-D. formulated the problem, implemented the grasping score, and designed the initial system and robot setup. I.T. designed and constructed most parts of the setup, including the integration of the tactile sensors, the routing of robot cables, and the fixtures for objects placements. A.R. supervised the project and the integration of its components and provided feedback on the manuscript. **Competing interests:** M.B. is now a research scientist at DeepMind, N.C.-D. is a Tech Lead at Samsung Research, I.T. is a staff research engineer at Boston Dynamics, and A.R. is affiliated with Boston Dynamics. **Data availability:** Videos of experiments as well as the object models, code, and data are posted at http://mcube.mit.edu/research/simPLE.html. See also the repository https://doi.org/10.5061/dryad.vdncjsz3q.

Submitted 13 June 2023
Accepted 29 May 2024
Published 26 June 2024
10.1126/scirobotics.adi8808

Science Robotics

SimPLE, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects

Maria Bauza, Antonia Bronars, Yifan Hou, Ian Taylor, Nikhil Chavan-Dafle, and Alberto Rodriguez

Sci. Robot. **9** (91), eadi8808. DOI: 10.1126/scirobotics.adi8808

Editor's summary
In robotic manipulation, there is often a trade-off between high accuracy for a repetitive motion and reliability in an unstructured environment. To teach a robot to move objects into an organized arrangement, Bauza *et al.* have developed a framework called SimPLE, which stands for Simulation to Pick, Localize, and placE. Given only a model of the object, the framework generates training data by sampling grasps in simulation. The SimPLE framework was tested with a set of 15 objects of different geometries on a dual-arm robot equipped with tactile sensors and an external depth camera. Using hand-to-hand regrasps, the robot successfully relocated the objects into structured arrangements, demonstrating the possibility of transferring a model learned in simulation to a real robot. —Melisa Yashinski

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.