

Navigating to Objects in the Real World

Theophile Gervet,^{1*} Soumith Chintala,⁴
Dhruv Batra,^{3,4} Jitendra Malik,^{2,4} Devendra Singh Chaplot⁴

¹Carnegie Mellon University,

²University of California, Berkeley,

³Georgia Institute of Technology,

⁴Meta AI Research

*To whom correspondence should be addressed; E-mail: tgervet@andrew.cmu.edu.

[Project Website](#)

Semantic navigation is necessary to deploy mobile robots in uncontrolled environments like our homes, schools, and hospitals. Many learning-based approaches have been proposed in response to the lack of semantic understanding of the classical pipeline for spatial navigation, which builds a geometric map using depth sensors and plans to reach point goals. Broadly, end-to-end learning approaches reactively map sensor inputs to actions with deep neural networks, while modular learning approaches enrich the classical pipeline with learning-based semantic sensing and exploration. But learned visual navigation policies have predominantly been evaluated in simulation. How well do different classes of methods work on a robot? We present a large-scale empirical study of semantic visual navigation methods comparing representative methods from classical, modular, and end-to-end learning approaches across six homes with no prior experience, maps, or instrumentation. We find that modular learning works well in the real world, attaining a 90% success rate. In contrast, end-to-end learning does not, dropping from 77% simulation to 23% real-world success rate due to a large image domain gap between simulation and reality. For practitioners, we show that modular learning is a reliable approach to navigate to objects: modularity and abstraction in policy design enable Sim-to-Real transfer. For researchers, we identify two key issues that prevent today’s simulators from being reliable evaluation benchmarks — (A) a large Sim-to-Real gap in images and (B) a disconnect between simulation and real-world error modes — and propose concrete steps forward.

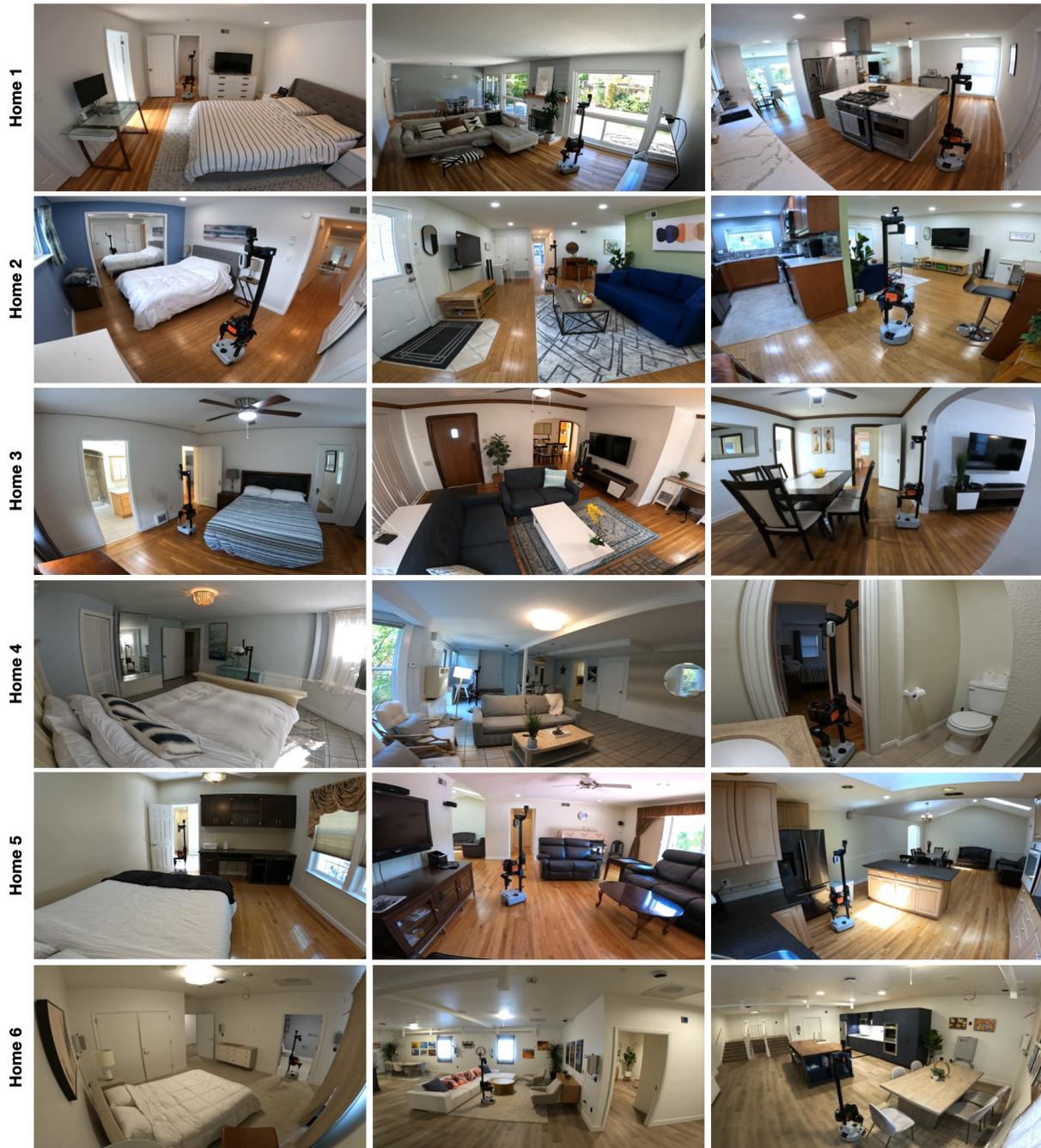


Fig. 1. Deployment of the semantic navigation policies in six visually diverse homes.

1 Introduction

Humans can navigate in unseen environments effortlessly. We can utilize our experience in prior environments to explore any new environment and find any target object efficiently. For example, when looking for a glass of water at a friend’s house we’re visiting for the first time, we can easily find the kitchen without going to bedrooms or storage closets. Learning such semantic priors is essential to deploy autonomous mobile robots in uncontrolled environments like our homes, schools, and hospitals. In this work, we tackle the Object Goal navigation [1] task where a robot is asked to find an object belonging to a particular category, like a bed or a couch, in a completely unseen environment.

Navigation has been studied in robotics literature for over three decades [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Many classical approaches to navigation require access to pre-computed maps [7, 10, 11]. Among approaches that can operate in unseen environments, classical approaches typically build a geometric map of the environment using depth sensors [12, 13, 14], and later a monocular RGB camera [15, 16, 17], while simultaneously localizing the robot relative to its growing map. Building on this simultaneous localization and mapping (SLAM) module, classical methods explore with heuristics such as frontier-based exploration [18] and leverage an analytical planner for low-level control to avoid obstacles and reach exploration or point goals. Adapting these methods to navigate to objects requires detecting objects, keeping objects in memory, and exploring semantically towards objects. Semantic SLAM methods [19, 20, 21, 22, 23, 24, 25] naturally extend SLAM to detect objects and keep them in memory but offer no solution for efficient semantic exploration.

Following recent advances in machine learning and computer vision, there has been a lot of interest in designing learning-based policies for visual navigation capable of learning these semantic priors. The most common learning-based methods for semantic navigation use a deep neural network, usually consisting of a visual encoder followed by a recurrent layer for memory, to predict actions directly from raw observations. These *end-to-end learning* approaches are trained using backpropagation with imitation learning (IL) or reinforcement learning (RL) losses. Inspired by seminal proofs of concept in autonomous driving [26, 27] and early successes of pixel-to-action deep reinforcement learning [28, 29], early applications of end-to-end learning to semantic navigation include [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. Most relevant to us, [41, 42, 43] propose end-to-end policies to navigate to object goals. End-to-end RL methods have been scaled to train with billions of frames corresponding to 80 years of navigation time (in sim) using distributed training [44] or tens of thousands of procedurally generated scenes [45]. While end-to-end policies often directly map sensor data to actions, akin to a modern version of Rodney Brook’s Subsumption architecture [46], researchers have also introduced some structure into the neural network, such as intermediate spatial [47, 48, 49, 50, 51] and topological representations [52, 53, 54].

Classical and end-to-end learning-based approaches offer distinct advantages. End-to-end learning policies can learn semantic priors for goal-directed exploration, while the modularity of the classical pipeline offers easier engineering and interpretability [55]. Following suc-

successful applications across robotics in autonomous driving [55, 56], flight [57], and grasping [58, 59, 60], much work has focused on designing *modular learning* approaches that aim to combine the benefits of both learning and classical methods. Modular learning approaches preserve the structure of the classical pipeline and replace analytical modules for specific sub-tasks with learned ones. In semantic navigation, the subtask decomposition typically includes separate modules for perception (object detection, mapping, pose estimation, SLAM), encoding goals, global waypoint selection policies, planning, and local obstacle avoidance policies. Learned modules are trained using direct supervision, which offers better sample efficiency than end-to-end learning [55]. Modular learning can enable Sim-to-Real transfer [56, 61, 62]: one can shield learned modules from Sim-to-Real domain gaps by designing abstractions of the input raw sensor data that contain sufficient information to solve the task while being invariant to environmental factors that are hard to simulate accurately (like photo-realistic RGB images) and thus unlikely to transfer from sim to reality. Representative examples of modular learning for exploration [61, 63] and for reaching object goals and image goals are [62, 64, 65, 66]. Most relevant to us, [62] cleanly isolates the problem of learning a policy to explore semantically towards objects from the rest of the navigation problem. Modular learning methods have also been applied to longer-horizon tasks such as following language navigation instructions [67, 68], executing language instructions interactively in ALFRED [69, 70, 71], object rearrangement in AI2 Thor [72, 73] and improving perception using active exploration [74, 75].

Over the past few years, the semantic navigation community has proposed hundreds of methods and organized tens of benchmarks in sim [76]. If the field doesn't lack proposed methods, what is missing to enable robots to navigate semantically? In our view, the missing piece of the puzzle is large-scale real-world evaluation. Learned navigation policies have predominantly been evaluated in sim [77]. We can attribute this to (A) the emergence of sophisticated embodied simulators [78, 79, 35], which significantly reduced the field's barrier to entry and sped up the proposal of new methods, and (B) the relative operational difficulties of bringing a robot out of the lab into diverse, realistic, deployment environments. But while simulators can be very useful for training, they are insufficient for evaluation. In the end, how well a method performs on a robot is the only thing that matters, and we have only partial answers, if any, as to whether sim is a good evaluation benchmark for semantic navigation. Does sim performance reflect real-world performance? Do design choices that improve sim performance improve real-world performance? Do sim error modes reflect real-world error modes? Most prior Sim-to-Real studies in navigation focused on spatial (point goal) navigation and legged locomotion [80, 81, 82, 83, 84], as opposed to semantic navigation. The few other real-world semantic navigation works directly train on real-world images for outdoor visual goals [85, 86] and language instruction following [87]. A lack of real-world evaluation opens semantic navigation to the risk of sim-only research that does not generalize to the real world [88].

Our proposed work addresses this issue through a large-scale empirical evaluation of semantic navigation policies. We compare representative methods from classical, modular learning, and end-to-end learning approaches across six visually diverse real home environments, as illustrated in Fig. 1. This represents 45 hours of robot experiments (3 methods x 6 homes x 10

episodes per home x 15 minutes per episode). In addition, we replicate one home in sim to ensure our experimental setting matches that of sim benchmarks and decompose the Sim-to-Real performance gap. We find that modular learning transfers to the real world very well, with performance rising from a 81% success rate in sim to 90% in the real world (within a limited time budget). In contrast, end-to-end learning performance drops sharply from 77% sim to 23% real-world success rate due to a large image domain gap between sim and reality. Classical approaches fall in between, with an 80% real-world success rate.

The takeaway for practitioners looking to build robots that navigate to objects is that the modular learning pipeline is very reliable, with a 90% success rate in limited time and efficient object search. In addition, we show that the remaining errors are primarily due to depth sensor failures (mapping failures and reflections in mirrors and TVs), which offers a clear path towards even greater reliability through better sensing or methods better able to deal with depth noise.

The takeaway for researchers is that much work remains to close the Sim-to-Real gap for semantic navigation. We identify two key issues that prevent today’s simulators, 3D assets, and task definitions from being reliable evaluation benchmarks. Then we propose concrete steps along two orthogonal paths forward: improve sim to better reflect real-world conditions and improve practices to work with imperfect sim.

First and foremost, there is a large Sim-to-Real gap in RGB images between sim and reality. Because of this, design choices easily overfit to sim. Two representative examples are that (A) policy architectures that directly operate on RGB images don’t transfer because they overfit to sim images, and (B) the common practice of training segmentation models on sim data improves performance in sim but hurts in the real world. Improving sim to close this gap would involve increasing RGB photo-realism or providing extensive plug-and-play RGB randomization, both hard open problems [88]. This leaves us with no choice but to improve practices to work with today’s sim. We should prioritize real-world transfer when designing policies: (A) replace policy architectures that directly operate on RGB images with ones leveraging abstractions as is common practice in other domains [56, 59, 89], and (B) avoid training a segmentation model on sim data if the policy architecture does not allow easily swapping it for one trained on real-world data at inference time.

Second, sim error modes don’t accurately reflect real-world error modes, which limits the usefulness of sim to diagnose bottlenecks and further improve methods. Modular learning errors in the real world largely stem from depth sensor errors, while sim benchmarks usually assume perfect depth or don’t provide realistic depth noise models. In contrast, errors in sim largely stem from reconstruction errors that do not happen in reality — both visual (imperfect RGB reconstruction that makes semantic segmentation harder in sim than reality) and physical (noisy navigation meshes that make planning harder in sim than reality). This explains the increase in performance from sim to reality for modular learning and stresses the need to always evaluate semantic navigation policies on real robots. We propose concrete steps forward to close this gap: (A) introduce realistic depth noise models for target deployment robots in sim benchmarks, (B) improve the visual quality of sim 3D scans, (C) improve the quality of sim navigation meshes. We hope our analysis sparks work to close the Sim-to-Real gap for semantic navigation.

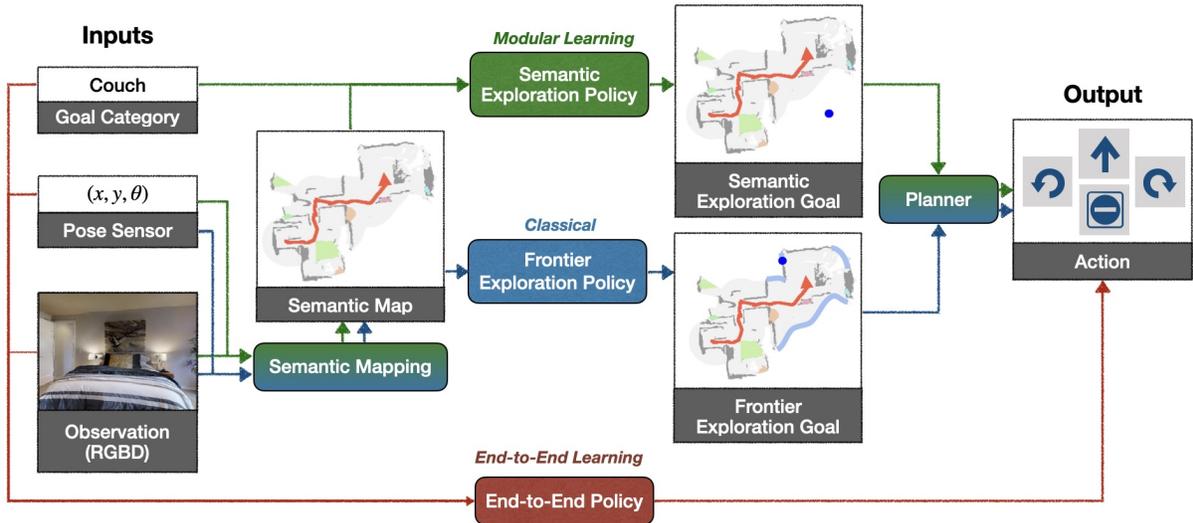


Fig. 2. Three approaches to navigate to objects. (A) The modular learning approach builds a top-down semantic map, selects a semantic exploration goal in this space, and plans low-level actions to reach this goal. (B) The classical approach also builds a semantic map but selects the closest unexplored region as the exploration goal, independently of the goal object. (C) The end-to-end learning approach directly maps sensor inputs and the goal object to low-level actions with a deep neural network.

One-Sentence Summary: Real-world empirical study of robot navigation methods comparing classical, end-to-end and modular learning approaches.

2 Results

[Movie 1](#) summarizes our results. We have deployed a policy representative of each of the classical, end-to-end learning, and modular learning approaches to Object Goal navigation on a Hello Robot Stretch robot [90]. Stretch is a lightweight, compact, low-cost mobile manipulator with an RGB-D camera and LiDAR, which we use only for localization and collision avoidance. We evaluated the policies at scale over 60 episodes split across six goal object categories in six different homes and a controlled study with one home replicated in sim. Before presenting our results, we give minimal formal background on the Object Goal task and the methods we evaluate.

2.1 Navigating to Objects: Task and Approaches

We consider the problem of semantic navigation instantiated by the Object Goal task [62, 1, 91]. In the Object Goal task, the robot’s objective is to navigate to an instance of a particular object category (in our case, “chair”, “couch”, “potted plant”, “toilet”, “tv”, or “bed”) as efficiently

as possible. The robot starts at a random location in an unknown home and receives the goal object category. At each step, the robot observes first-person RGB and depth images and takes a discrete navigation action: move forward (25 cm), turn left or right (30 degrees), or stop. The robot needs to take the stop action when it believes it has reached the goal object. An episode is considered successful if the robot’s distance to an instance of the goal object category is less than some threshold (1 meter) when the robot takes the stop action. In contrast, an episode is a failure if the agent calls the stop action too far from the goal object, never calls the stop action before a fixed maximum number of timesteps (500), or collides too many times (more than 20) with its environment. We use two metrics for comparing methods: Success Rate (SR), the ratio of successful episodes, and Success weighted by Path Length (SPL) [1], the ratio of path length over optimal path length for successful episodes, which measures exploration efficiency. The Object Goal task requires spatial scene understanding (obstacle and navigable space detection), semantic scene understanding (object detection), learning semantic priors (for efficient exploration), and episodic memory (keeping track of explored and unexplored areas).

We evaluate three methods, each representative of one class of approaches. To represent modular learning for Object Goal navigation, we picked [62]. To represent classical approaches, we replace the semantic exploration policy of [62] with frontier exploration [18], which navigates towards the closest unexplored region. Finally, to represent end-to-end learning, we picked [43]. We will describe each method and what makes it representative of its class of approaches in detail in the Materials and Methods. For now, Fig. 2 illustrates all the necessary background to understand the Results and Discussion.

2.2 Natural Home Environments

We deployed navigation policies representative of each class of approaches in six natural home environments never seen before in simulation or reality. We evaluated the policies over 60 episodes split across the six goal object categories and six homes. This represents 45 hours of robot experiments (3 methods x 6 homes x 10 episodes per home x 15 minutes per episode).

Fig. 3 illustrates our main results quantitatively. Classical and modular learning approaches perform better in the real world than simulation, up from 78% to 80% and 81% to 90% success rate, respectively. In contrast, end-to-end learning performs much worse in the real world, down from 77% to 23% success rate. This trend holds across all homes and all goal object categories, as shown in Tables S1 and S2. Fig. 4 and S2 compare all approaches on the same episode to illustrate these results qualitatively. Fig. S3 illustrates two typical failure modes of the end-to-end learning policy, besides collisions. First, the policy often detects the goal object but fails to stop nearby, which is consistent with prior work observing this “last mile” failure in simulation [41]. Second, the policy revisits the same locations, often semantically unrelated to the goal object. These failure modes seem to display a lack of semantic understanding, a lack of long-term memory, and poor exploration. Fig. S1 and Movie 1 illustrate how the modular learning approach improves over the classical approach. The learned exploration policy leverages the semantics of the top-down map to search for a specific goal object effectively. In contrast,

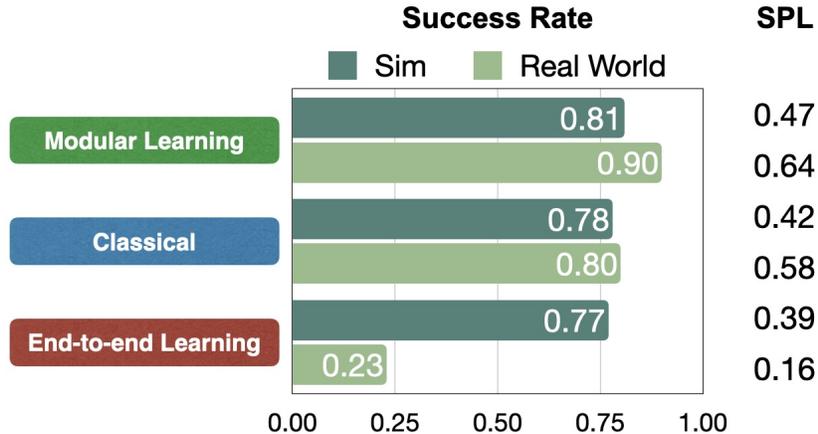


Fig. 3. Navigation performance in simulation vs. real at scale. We compare the Success Rate (SR) and Success weighted by Path Length (SPL) of methods representative of classical, end-to-end-learning, and modular learning approaches on large real world (60 episodes in 6 homes) and simulation datasets (single-floor navigation episodes of val split of the 2022 Habitat Challenge with 1093 episodes in 20 simulated homes of the HM3D Semantics dataset [92]). (A) Performance for all methods is comparable in simulation, at around 80% success rate. (B) Classical and modular learning approaches transfer well, up from 78% to 80% and 81% to 90%, respectively. (C) End-to-end learning fails to transfer, down from 77% to 23% success rate.

the frontier exploration policy, which selects the closest unexplored region as the exploration goal independently of the goal object, exhibits depth-first search behavior and fails to backtrack using semantics.

2.3 Controlled Experiments in a Home Replicated in Simulation

We ran a controlled study in one home replicated in simulation. This serves two purposes. First, this lets us decompose the discrepancy between results on a simulation benchmark and the real world into (A) the discrepancy between the sim benchmark and the sim replica and (B) the discrepancy between the sim replica and the real home. This lets us verify that our experimental setting matches the sim benchmark: results in our sim replica should be close to that of the sim benchmark. Second, this allows us to run an ablation study to select the best end-to-end policy to evaluate at scale across all six homes. We replicate a single home in simulation as 3D scanning and digitizing a house — with the same Matterport camera used to digitize homes of the sim benchmark [94] — takes a few hours.

We first present the results of the ablation study we conducted to select the best end-to-end policy to evaluate at scale across all six homes. Modular learning and classical approaches only use first-person RGB-D and predicted segmentation images through a top-down semantic map. This makes them independent of changes in camera parameters and lets us easily swap semantic segmentation models between training and evaluation. In contrast, end-to-end learning

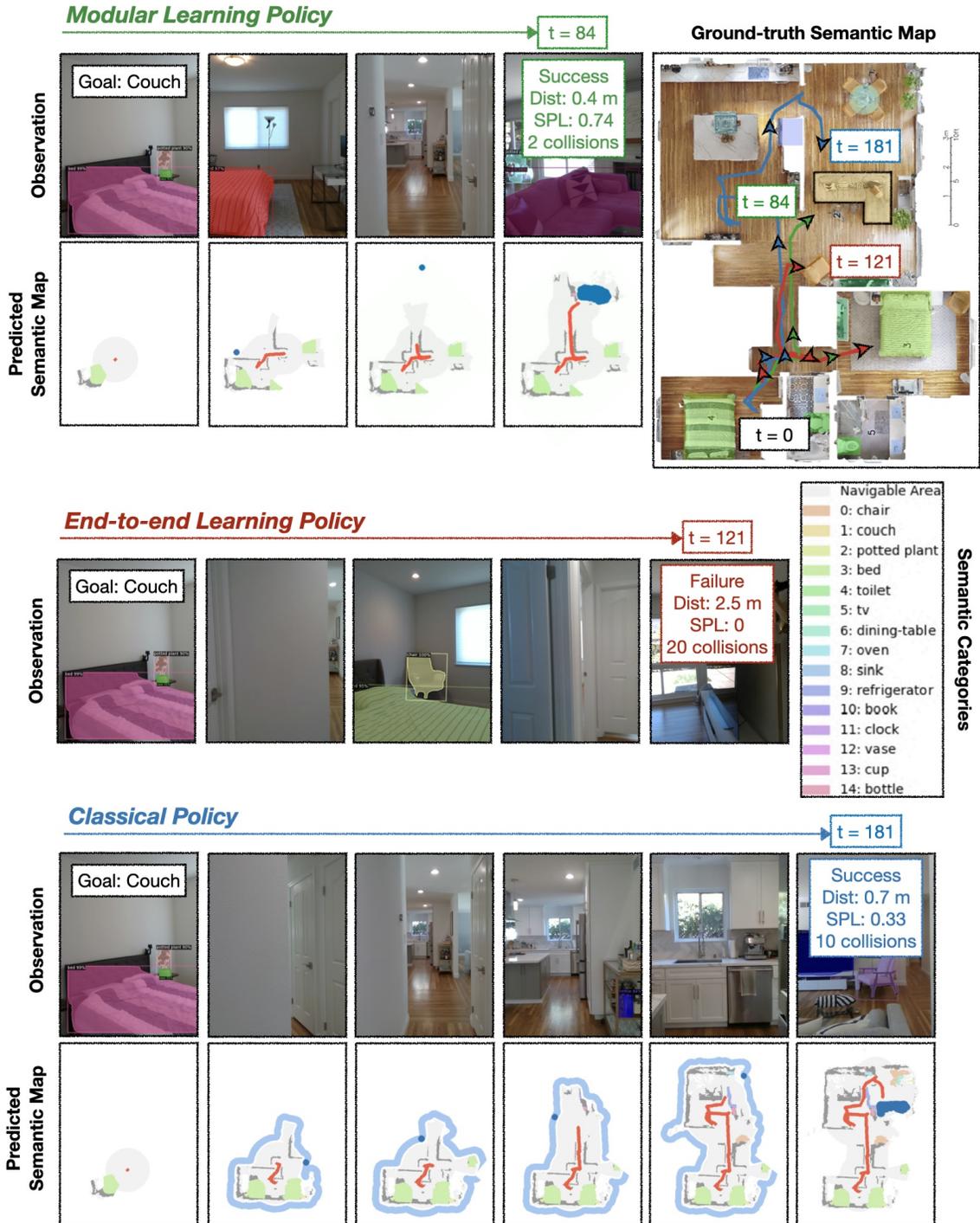


Fig. 4. Three approaches on the same episode. (A) Modular learning reaches the couch goal in 84 steps (SPL = 0.74). (B) End-to-end learning collides too many times (20 max) after 121 steps. (C) The classical policy reaches the goal after 181 steps and a detour through the kitchen (SPL = 0.33).

Table 1. Navigation performance in simulation vs. real on the same episodes in the same home.

We compare policies by Success Rate (SR) on a sim benchmark (single-floor navigation episodes of val split of the 2022 Habitat Challenge with 1093 episodes in 20 simulated homes of the HM3D Semantics dataset [92]), and the same home in sim and the real world (10 episodes). We compute the Sim-vs-Real Correlation Coefficient (SRCC) introduced in [93], which is the Pearson correlation between binary episode outcomes in sim vs. real (in $[0, 1]$, higher is better). **(A)** Top table: In an ablation study of end-to-end policies, we vary training camera parameters, the segmentation model training domain, and the policy training algorithm. Bottom table rows 1-4: Real-world performance of end-to-end policy ablations is inversely correlated to sim performance. Policy 1 is the best in sim but the worst in the real world, while policy 4 — which we selected for large-scale evaluation — is the worst in sim but the best in the real world. All variants other than policy 4 have a low SRCC, indicating their design overfits to sim. **(B)** Bottom table rows 5-7: Performance on the sim replica is close to that on the sim benchmark, which shows that our experimental setting matches that of the sim benchmark. Although, performance in the real-world home is close to that on its sim replica, the SRCC is still relatively low because policies fail different episodes in sim and the real world. As we will see in the Discussion, this is due to a disconnect between error modes in sim and reality.

End-to-end learning ablations

End-to-end Policy	Policy training settings		
	Camera parameters	Segmentation training domain	Policy training algorithm
End-to-end 1	Sim Benchmark	Sim	IL + RL
End-to-end 2	Robot	Sim	IL + RL
End-to-end 3	Robot	Real-world	IL + RL
End-to-end 4	Robot	Real-world	IL

Evaluation results

Navigation policy	Sim benchmark SR	Real home sim replica SR	Real home SR	Real home SRCC
End-to-end 1	0.77	0.80	0.00	0.20
End-to-end 2	0.71	0.70	0.00	0.30
End-to-end 3	0.61	0.60	0.10	0.40
End-to-end 4	0.48	0.50	0.30	0.60
End-to-end	0.48	0.50	0.30	0.60
Modular	0.81	0.80	0.90	0.70
Classical	0.78	0.80	0.90	0.70

approaches directly operate on RGB-D and predicted segmentation images. This means varying camera parameters between training and evaluation settings introduces a domain gap, and an end-to-end policy is closely tied to the segmentation model used in its network architecture during training. These characteristics of end-to-end policies mean that in order to give them the best chance to work on a robot, an ablation study is needed to select the best-performing policy.

Table 1 (A) shows results of this ablation study. We train four different end-to-end policies, varying the camera parameters, segmentation model training domain, and training algorithm. We measure each policy’s performance in sim, both on a large-scale benchmark and our sim replica, and in the real home. Overall, Policy 1, trained with the camera parameters of the sim benchmark, which we detail in the Materials and Methods section, and using a segmentation model trained in sim, performs the best in sim but the worst in the real world. Policy 2, trained with robot camera parameters, performs slightly worse in sim, which shows that our robot camera parameters make the task slightly harder than the sim benchmark camera parameters. Policies 3 and 4 that use a segmentation model trained in the real world perform worse in sim but better in the real world. This shows that using a segmentation model trained in sim is overfitting to sim. Policy 4, which is trained with imitation learning (IL) only as opposed to IL followed by reinforcement learning (RL) fine-tuning, performs the worst in sim but the best in reality. This shows that RL fine-tuning can overfit to sim. We select policy 4 to evaluate at scale across all six homes.

Overall, these results show that performance in sim is often inversely proportional to performance in the real world because design choices to improve performance in sim can easily overfit to sim. We can quantify this observation through the Sim-vs-Real Correlation Coefficient (SRCC) introduced in [93], which measures how well sim results can predict real-world results. The SRCC is low for all end-to-end policy variants, starting as low as 0.20 for policy 1, which performs best in sim, and increasing to 0.60 for policy 4 as we correct for design choices that overfit to sim.

Now that we have selected the best end-to-end policy to evaluate at scale, Table 1 (B) decomposes the discrepancy between results for all selected methods on the sim benchmark and the real world into the discrepancy between the sim benchmark and the sim replica, and the discrepancy between the sim replica and the real home. First, performance on the sim replica is close to that on the sim benchmark, which verifies that our experimental setting matches that of the sim benchmark. Second, even though absolute performance is fairly close in sim and the real world for the policies we evaluate at scale (e.g., 0.80 sim vs. 0.90 real success rate for modular learning), the SRCC is still low (e.g., only 0.70 for modular learning) because each method fails different episodes in sim and the real world. As we will see in the Discussion section, this is due to a disconnect between error modes in sim and reality.

3 Discussion

In this section, we (A) analyze error modes of the modular learning approach in sim vs. real to make sense of its rise in performance from sim to real, (B) emphasize the significance of the 90% real-world success rate of modular learning, and (C) investigate why modular learning transfers better than end-to-end learning and reflect on what this means for the field.

3.1 Modular Learning Error Modes in Real-world vs. Simulation

To make sense of the rise from 81% sim to 90% real-world success rate for the modular learning approach, we compare its sim and real error modes in Table S3. Surprisingly, Table S3 shows there is nearly no overlap between sim and real error modes. Errors in the real world largely stem from depth sensor errors (5 out of 6 total errors), as illustrated in Fig. S4 and Movie 1, while the sim benchmark assumes perfect depth sensing. When approaching a door at an angle, noise in depth can block it in the map, making a room inaccessible without a map denoising mechanism. Reflection in mirrors and TVs can also cause depth sensor errors and downstream navigation failures. In contrast, a lot of the failures that occur in sim are due to reconstruction errors — both visual and physical — which do not happen in reality. Indeed, 10.1% out of the total 18.6% episode failures in the sim benchmark are due to segmentation errors, while segmentation errors did not cause any episode failures for modular learning in the real world. Segmentation errors are more common in sim because visual reconstruction can make objects unrecognizable, as illustrated in Fig. S5 (A) and Movie 1. Physical reconstruction errors represent another 5.5% of the total 18.6% episode failures in sim. They lead to noisy navigation meshes with narrow pathways that are hard to navigate for discrete planners that work well in the real world, as illustrated in Fig. S5 (B) and Movie 1. This gap in error modes explains the performance gap between sim and reality for modular learning.

The lack of overlap between sim and real-world error modes is concerning because it limits the usefulness of simulation to diagnose bottlenecks and further improve policies. This is a practical working definition of the Sim-to-Real gap: we only care about improving sim realism to the extent that it lets us develop better real-world policies. A gap in error modes prevents us from doing so. This stresses the need to always evaluate semantic navigation policies on real robots for results to be meaningful. Based on our analysis, we propose concrete steps forward to close this gap: (A) introduce realistic depth noise models for target deployment robots in sim benchmarks, (B) improve the visual quality of sim 3D scans, (C) improve the quality of sim navigation meshes.

3.2 Towards Solving Navigation to Objects with Modular Learning

The main takeaway of this study for practitioners looking to build robots that navigate to objects is that the modular learning pipeline is very reliable, with a 90% success rate in limited time and efficient object search with an SPL of 0.64. In addition, we show that the remaining errors are

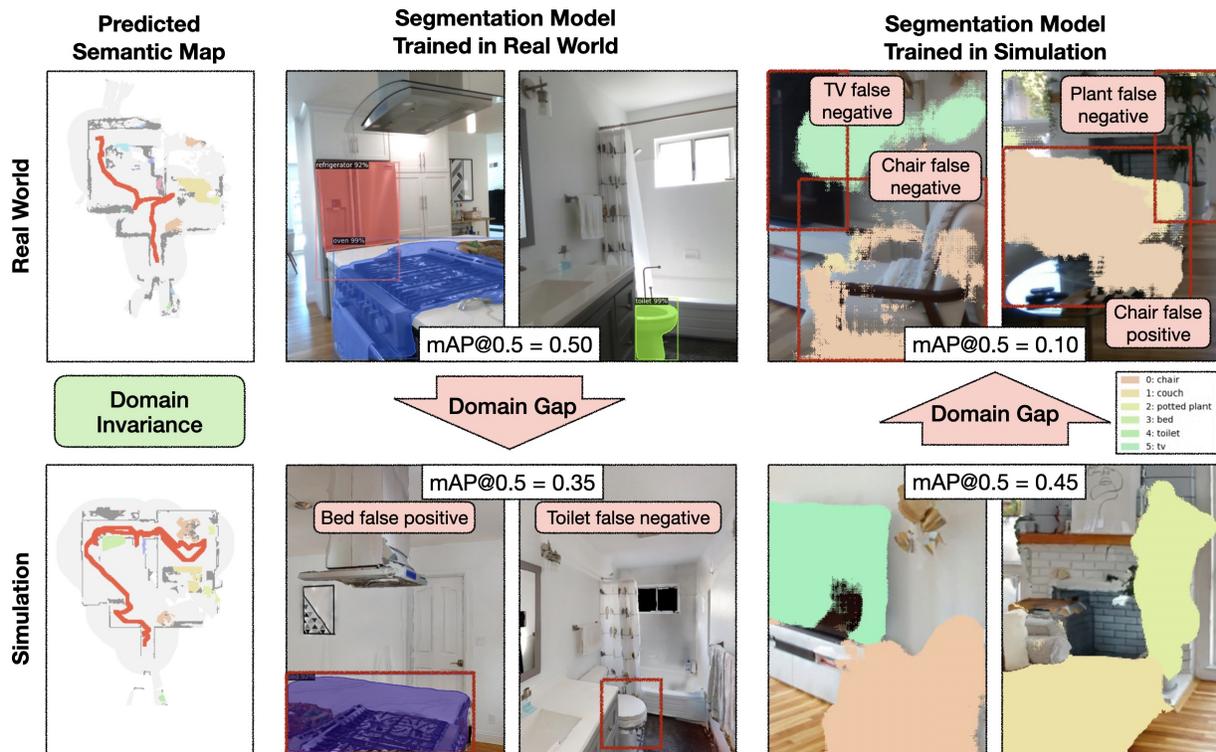


Fig. 5. Sim-vs-Real domain invariances, gaps, and their effects on segmentation. From left to right, all images come from episodes in our controlled study: (A) The semantic map space is invariant between the real world and simulation. (B) The image space exhibits a large gap between the real world and simulation. (C) This gap causes a large drop in performance when transferring a segmentation model trained in the real world to simulation and vice versa.

primarily due to depth sensor failures, which offers a clear path towards even greater reliability through better sensing or methods better able to deal with depth noise.

One limitation of our study is the restriction to six goal object categories to align with current sim benchmarks. All methods we evaluate are straightforward to extend to a larger finite set of categories, and opportunities for future work include extensions to an unbounded set of object categories with open-vocabulary detectors [95] and instance-level object goals [96].

3.3 Why Modular Learning Transfers Better than End-to-end Learning

Fig. 5 illustrates why modular learning transfers better than end-to-end learning. The semantic exploration policy of the modular learning approach takes a semantic map as input, while the end-to-end policy directly takes the RGB-D frames as input. The figure shows that the semantic map space is invariant between simulation and the real world. This is due to the semantic map abstracting away pixels in favor of semantic categories and reducing the spatial granularity of the map voxel size (5 cm in our experiments). In contrast, the RGB image space exhibits a large

gap between the real world and simulation because current reconstruction engines cannot yet generate photo-realistic images. This causes a large drift between the training and evaluation domains for the deep neural network of the end-to-end policy.

The significance of the gap from simulation images to real-world images is well illustrated by the segmentation errors it causes in Fig. 5. A segmentation model trained in the real world suffers a severe performance drop when transferred to simulation, down from 0.50 to 0.35 mean average precision at a 0.50 intersection over union threshold (mAP@0.5). Similarly, a segmentation model trained in simulation suffers an even larger performance drop when transferred to the real world (down from 0.45 to 0.10 mAP@0.5). If semantic segmentation transfers poorly from simulation to reality, it is reasonable to expect an end-to-end semantic navigation policy trained on sim images to transfer poorly to real-world images. While the image domain gap affects a segmentation model over a single prediction, the gap accumulates over many predictions made over a long horizon for an end-to-end navigation policy.

Because of this image domain gap, design choices easily overfit to simulation. Two representative examples in our results are that (A) end-to-end policy architectures that directly operate on RGB images don't transfer because they overfit to simulation images, and (B) the common practice of training segmentation models on simulation data improves performance in simulation but hurts in the real world.

How can we address this image domain gap? At this point, it is worth taking a step back to briefly survey how other domains of robotics tackle Sim-to-Real gaps. Prevalent options today [88] are: (A) train on real-world data, (B) train in simulation with domain randomization, and (C) train in simulation with modularity and abstraction. Training on real-world data bypasses any Sim-to-Real gap but is expensive, slow, and potentially unsafe. It is the option of choice for perception stacks across robotics and has been applied to end-to-end control in domains where sub-optimal policy operation is safe, like grasping and static manipulation [97, 98]. But it is not straightforward to scale up for semantic navigation as mobile robots require constant supervision. This leaves us with simulation training. Domain randomization [99] — randomizing environmental factors that are hard to simulate accurately during training to make policies robust to these factors — has been applied successfully in domains where randomization is straightforward like dexterous manipulation [100] where one needs to randomize over physics and single object appearance. But it is still an open problem whether RGB image randomization can be scaled up to entire houses to bring the same robustness to semantic navigation [88]. Our final option, training in simulation with modularity and abstraction [56] — designing abstractions of the input raw sensor data that contain sufficient information to solve the task while being invariant to environmental factors that are hard to simulate accurately — is the practice of choice in autonomous flight [101, 89], legged locomotion [84, 83], grasping [59, 58], and a promising path in autonomous driving [56].

In our view, training in simulation and Sim-to-Real transfer via modularity and abstraction is the most promising path forward for semantic navigation. Example semantic abstractions could be first-person semantic segmentation masks [56], topological scene representations [53], or top-down spatial semantic maps [62]. One limitation of our study and opportunity for future

work is that we didn’t take further steps in this direction — like evaluating an end-to-end policy abstracting RGB frames through semantic frames.

4 Materials and Methods

We deployed three navigation policies on a robot across six homes to study how well different methods to navigate to object categories transfer from simulation to the real world. This section outlines our Sim-to-Real transfer methodology, details the specific methods we evaluate, and explains what makes them representative of broader classes of approaches.

4.1 Sim-to-Real Transfer Methodology

Hardware and Software Stack: We deployed navigation policies on a Hello Robot Stretch robot [90]. We selected Stretch because it is low-cost, lightweight, and compact, allowing us to transport it out of the lab into real homes easily. We trained all learned components of navigation policies in simulation with the Habitat platform [102]. We selected Habitat for its simulation speed, which is crucial for training navigation policy components with reinforcement learning. At inference time, policies were deployed with the fairo library [103] to run the same navigation code in simulation and the real world.

Matching Policy Inputs and Outputs from Sim to Real: All navigation policies we evaluated were trained in simulation and had so far only been evaluated in simulation. To transfer these policies to a robot and be able to compare real-world results to simulation, we had to match the Object Goal task input and output spaces from sim on the robot. In the Object Goal task, the robot receives an RGB-D image and a sensor pose at each step. Stretch’s Intel RealSense D435i camera gives us (640 x 480) RGB-D images with a 42-degree horizontal field of view. We preprocess the depth with standard spatial and temporal filters and threshold all values beyond the camera’s confidence range (4 meters). We use off-the-shelf LiDAR-based Hector SLAM [104] to estimate the robot’s pose. We also use LiDAR for collision detection to deploy policies safely. To guarantee a fair comparison to simulation settings, we restrict the use of LiDAR to collision detection and localization and do not use it for mapping. Given the output action space is discrete — move forward 25 cm, turn left/right 30 degrees, or stop — it is straightforward to match on the robot. Sim-to-Real transfer is illustrated in Fig. 6 (A).

Generating Episodes in Real Homes: We evaluate the robot on 60 episodes across six object goal categories (“chair”, “couch”, “potted plant”, “toilet”, “tv”, and “bed”) in six homes. We selected these goal object categories to be able to directly compare results in the real world with that of a leading sim benchmark: the Habitat 2022 Object Navigation Challenge. We selected homes for their visual diversity and size. To generate an episode within a home, we selected a goal category and a starting location for the robot while trying to balance the distribution of goal object categories and match the distribution of geodesic distances to the goal to that of the Habitat Challenge, as shown in Fig. S6. We consider an episode successful if the robot called

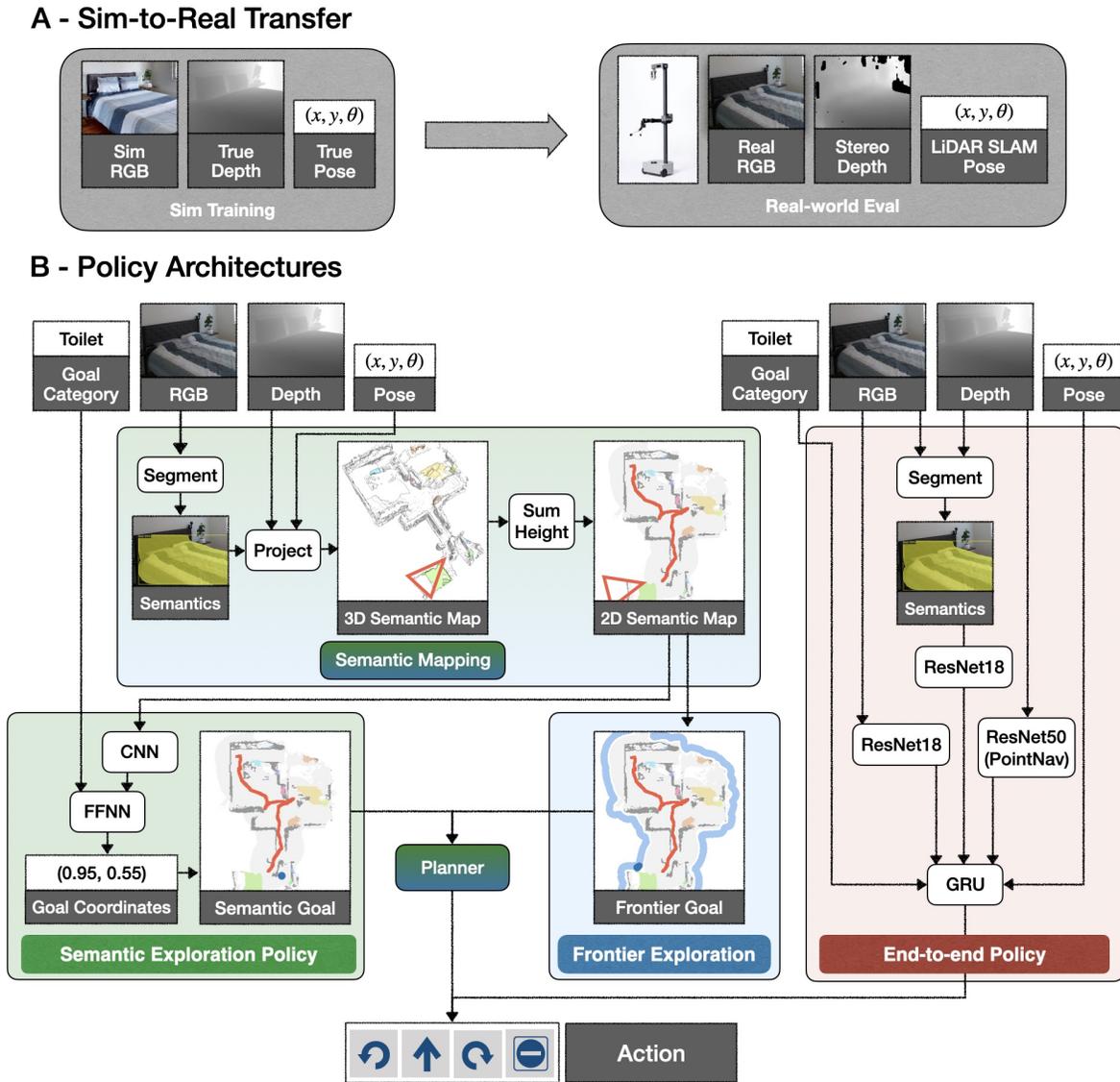


Fig. 6. Three approaches to navigate to objects and their Sim-to-Real transfer. (A) We match policy inputs from sim in the real world: we estimate depth with a stereo camera, and the pose with LiDAR-based SLAM. (B) Policy architectures in detail from left to right, top to bottom. The semantic mapping module of classical and modular learning approaches segments the RGB frame, projects it into a 3D semantic voxel map using the depth and pose, and sums over the height to compute a 2D semantic map. The semantic exploration policy of the modular learning approach predicts an exploration goal with a feed-forward neural network (FFNN) as a function of map features computed with a convolutional neural network (CNN) and the goal object. The frontier exploration policy of the classical approach explores the closest unexplored region independently of the goal object. Both approaches plan low-level actions to reach the high-level exploration goal. The end-to-end approach segments the RGB frame with a pre-trained segmentation model, computes RGB, depth, and semantic features with CNN backbones, which can be also be pre-trained, and directly predicts low-level actions from frame features, the goal object, and the pose with a Gated Recurrent Unit (GRU).

the stop action close enough (less than 1 meter) to an instance of the goal category within the time (200 steps) and collision (20 collisions) budgets. To compute the Success weighted by Path Length (SPL) [1], we measure the geodesic distance to the goal object instance closest to the starting location. To replicate one home and its episodes in simulation in our controlled study, we scanned and digitized the home with a Matterport Pro2 3D camera and matched the real-world goals and starting positions in sim.

Fig. 6 (B) illustrates the architectures of the three policies we evaluate. We first describe the classical and modular learning semantic mapping approaches before turning our attention to end-to-end learning approaches.

4.2 Classical and Modular Learning Approaches

Extending Classical Approaches to Navigate to Objects: Classical simultaneous localization and mapping (SLAM) approaches to spatial navigation typically build a spatial map of the environment while simultaneously localizing the robot relative to its growing map [12, 13], and navigate to geometric points within this map via path planning. Adapting these methods to navigate to objects requires detecting objects, keeping objects in memory, and exploring semantically towards objects. Semantic SLAM methods [19, 20, 21, 22, 23, 24] naturally extend SLAM to detect objects and keep them in memory in a spatial semantic map but offer no solution for efficient semantic exploration. Among the many heuristics for goal-agnostic exploration proposed in the classical navigation literature, we select frontier-based exploration [18] — navigate towards the closest unexplored region — to represent classical exploration methods because it was particularly effective in prior work [62]. As shown in Fig. 6 (B), this completes a pipeline representative of classical approaches: semantic mapping to keep seen objects in memory, frontier exploration to select high-level goals, and path planning to select low-level actions.

Note that if we only care about navigating to a single object starting with zero information about the environment — the setting we evaluate in this paper — we don’t even need to keep objects in memory in a semantic map in the classical approach. We can simply use a geometric map for exploration and planning and head toward the goal object as soon as we detect it in the frame with a pre-trained object detector. We add semantic mapping to the classical approach for ease of implementation — we can head towards an object by projecting it into a single semantic channel in the map and planning toward it — and to make it applicable to the more practical but harder-to-evaluate setting where the robot might need to execute several such object search commands in a row.

Modular Learning to Navigate to Objects: While frontier exploration is effective, as shown in our results, it is necessarily suboptimal because it ignores the goal object. We intuitively understand where a “couch” is more likely to be found. This is where modular learning comes into play: can we train an exploration policy to leverage the statistical regularities in the layout of objects in homes to explore more efficiently? This is what the semantic exploration method

proposed in [62] does. We selected it to represent modular learning as it cleanly isolates the problem of learning an exploration module from the rest of the navigation problem. As shown in Fig. 6 (A), it preserves the structure of the classical pipeline to navigate to objects presented above but replaces frontier exploration with a learned semantic exploration module. With this high-level picture in mind, we explain each component in detail.

Semantic Map Representation: The semantic map is a spatial representation of the environment that keeps track of objects, obstacles, and explored areas. Concretely, it is a binary $K \times M \times M$ matrix where $M \times M$ is the map size and K is the number of map channels. Each cell of this spatial map corresponds to 25 cm^2 ($5 \text{ cm} \times 5 \text{ cm}$) in the physical world. Map channels $K = C + 4$ where C is the number of semantic object categories, and the remaining 4 channels represent the obstacles, the explored area, and the agent’s current and past locations. An entry in the map is one if the cell contains an object of a particular semantic category, an obstacle, or is explored, depending on the channel, and zero otherwise. The map is initialized with all zeros at the beginning of an episode and the agent starts at the center of the map facing east.

Semantic Mapping Module: In order to build the semantic map, we need to predict semantic categories and segmentation masks of objects in first-person observations. We use a Mask-RCNN [105] with ResNet50 [106] backbone pretrained on MS-COCO for object detection and instance segmentation. We project first-persons semantic segmentation into a point cloud using the depth, bin the point cloud into a 3D semantic voxel map, transform it from the robot’s coordinate system to that of the semantic map using the robot pose, and finally sum over the height to compute a 2D semantic map.

Frontier Exploration Policy: With our semantic map at hand, it is straightforward to implement frontier exploration: each step, we select the boundary between the explored and unexplored region of the map, i.e., the frontier, and within this boundary select the point closest to the robot in geodesic distance. This exploration strategy exhibits depth-first search behavior: once the robot heads into a direction, the closest unexplored region is in front of it until an obstacle blocks the way. As soon as the goal object is seen, i.e., as soon as the channel of the semantic map for the object goal has a nonzero element, we stop exploring and select all non-zero elements as the goal.

Semantic Exploration Policy: The semantic exploration strategy decides a goal as a function of the current semantic map and the goal object. This requires learning semantic priors on the relative arrangement of objects and areas in homes. As shown in Fig. 6 (B), map features are computed with a convolutional neural network (CNN) and passed through a feed-forward neural network (FFNN) along with a learnable embedding for the goal object to compute a goal in $[0, 1]^2$, which is then converted to top-down map space. The policy is trained using reinforcement learning (RL) with the distance reduced to the nearest goal object as the reward. We use a policy trained in [62] on home layouts from the Gibson dataset [107]. As in [62], we sample the long-term goal at a coarse time-scale, once every 25 steps. This reduces the time-horizon for exploration in RL exponentially and consequently, reduces the sample complexity.

As for the frontier exploration policy, as soon as the goal object is seen, we stop exploring and select it as the goal.

Planner: Given a long-term goal output by the frontier or semantic exploration policy, we use the Fast Marching Method [108] as in [62] to plan a path and the first low-level action along this path deterministically. Although the semantic exploration policy acts at a coarse time scale, the planner acts at a fine time scale: every step we update the map and replan the path to the long-term goal.

Sim-to-Real Transfer: Semantic mapping methods are straightforward to transfer to the robot as they are independent of camera parameters and the semantic map space is invariant between sim and the real-world, as presented in the Discussion section.

4.3 End-to-end Learning Approaches

End-to-end Learning to Navigate to Objects: In contrast to modular approaches that explicitly build a semantic map of the environment and plan actions, end-to-end learning approaches directly predict actions from raw sensor data with a deep neural network. The network needs to learn to understand the scene spatially and semantically, keep track of long-term memory, and plan actions. Given each of these tasks is hard in isolation, end-to-end learning approaches typically require tens of thousands of expert demonstrations or hundreds of millions of steps of reinforcement learning to train. Given neither of these can realistically be collected in the real world for our Object Goal navigation task, training must be done in simulation. As illustrated in Fig. 6 (B), at each step, the typical end-to-end architecture for long horizon tasks with high-dimensional image input computes image features with one or multiple convolutional neural networks, which can be trained from scratch or pre-trained, and keeps track of memory and plans implicitly with a recurrent neural network, in our case a gated recurrent unit (GRU) [109].

Habitat-Web Policy: We selected Habitat-Web [43] as a representative example of end-to-end learning for the Object Goal navigation task because it closely follows the above paradigm and has the highest performance on a leading sim benchmark, the 2022 Habitat Challenge, at the time of publication. Habitat-Web [43] trains an end-to-end policy from human demonstrations on the Object Goal task in the Habitat simulator. We preserve the general architecture of the policy and swap different components in our ablation study to find the policy that works best in the real world. Semantic segmentation is predicted from RGB (and possibly depth) with a pre-trained and frozen segmentation model. First-person RGB, depth, and semantic frame features are computed with multiple CNN backbones, leveraging pre-trained models when available to reduce sample complexity. RGB, depth, and semantic frames are then fed into a GRU [109] along with other low dimensional features — the robot pose, a learnable goal object embedding, the fraction of the visual input occupied by the goal category, and the previous action — to predict a distribution over next actions.

Architecture and Training Details: The segmentation model used in the original architecture

is a RedNet [110] pre-trained on the SUN RGB-D dataset [111] and fine-tuned on images from the Habitat simulator. As shown in our results, using a segmentation model fine-tuned in simulation hurts performance in the real world. In our ablation study of end-to-end architectures, we thus replace this model with the same Mask-RCNN pre-trained on MS-COCO used by modular semantic mapping methods. Depth features are computed with a ResNet50 pre-trained on Point Goal navigation [112]. RGB and semantic frame features are computed with a ResNet18 trained from scratch. The network is first trained with imitation learning (IL) on 80,000 human demonstrations (or approximately 20 million actions) collected for the Object Goal task in the Habitat simulator over three days with 128 Nvidia V100 32GB GPUs. It is then fine-tuned with reinforcement learning (RL) with a sparse binary reward when the goal is found, over 150 million steps in an additional three days on 32 Nvidia V100 32GB GPUs. Our ablation study of end-to-end policies shows that RL fine-tuning helps in sim but hurts in the real world.

Sim-to-Real Transfer: While semantic mapping methods are independent of camera parameters, end-to-end methods directly operate on first-person frames, and any change in camera parameters causes a large domain gap. Given our robot camera parameters are different from those used to train the original Habitat-Web policy — (640 x 480) frames with a 42° horizontal field of view, as opposed to (480 x 640) frames with a 79° horizontal field of view — we retrain the policy with robot camera parameters replicated in simulation.

As shown in Fig. 6 (A), real-world depth estimated via stereo camera is noisy, which introduces an additional domain gap. To compensate for this, we tried training with the indoor depth noise model provided by Habitat [113], as we didn’t have any noise model tuned specifically for our robot’s Intel RealSense D435i camera. We found this to hurt real-world performance, suggesting this noise model doesn’t match our robot’s camera noise.

While we can easily swap a different segmentation model in modular semantic mapping methods, changing the segmentation model causes a large domain gap for end-to-end methods and requires training a new policy.

In contrast to simulation, robot discrete actions are also not deterministic in the real world due to actuation noise: a 25 cm forward move or 30 degree turn command will not have the exact intended effect. The usual way to compensate for this is to simulate actuation noise during training. We did not simulate actuation noise because a prior Sim-to-Real study of Point Goal navigation [93] found this to hurt real-world performance.

Finally, while we can easily inspect the output of various modules — like the semantic map, the exploration goals, or the plans — to diagnose issues when transferring modular methods, our only recourse when transferring end-to-end policies is to try matching training and evaluation inputs as closely as possible. All these considerations make transferring end-to-end policies much more challenging than modular methods.

5 Acknowledgments

We thank Saurabh Gupta for reviewing the manuscript and Brandon Trabucco for lending his voice to the video.

6 Supplementary Materials

Fig. [S1](#). Comparison of modular learning and classical policies.

Fig. [S2](#). Three approaches on the same episode.

Fig. [S3](#). Failure modes of end-to-end learning.

Fig. [S4](#). Real-world depth sensor error modes.

Fig. [S5](#). Simulation visual and physical reconstruction error modes.

Fig. [S6](#). Distributions of goal objects and geodesic distances to goal in reality vs. sim.

Table [S1](#). Navigation performance on all 60 episodes.

Table [S2](#). Navigation performance aggregated by home and goal object.

Table [S3](#). Modular learning real vs. sim error modes.

Table [S4](#). Modular learning programmatic error analysis on sim benchmark.

Table [S5](#). Modular learning manual analysis of remaining errors on sim benchmark.

Table [S6](#). Modular learning episode outcomes on sim benchmark.

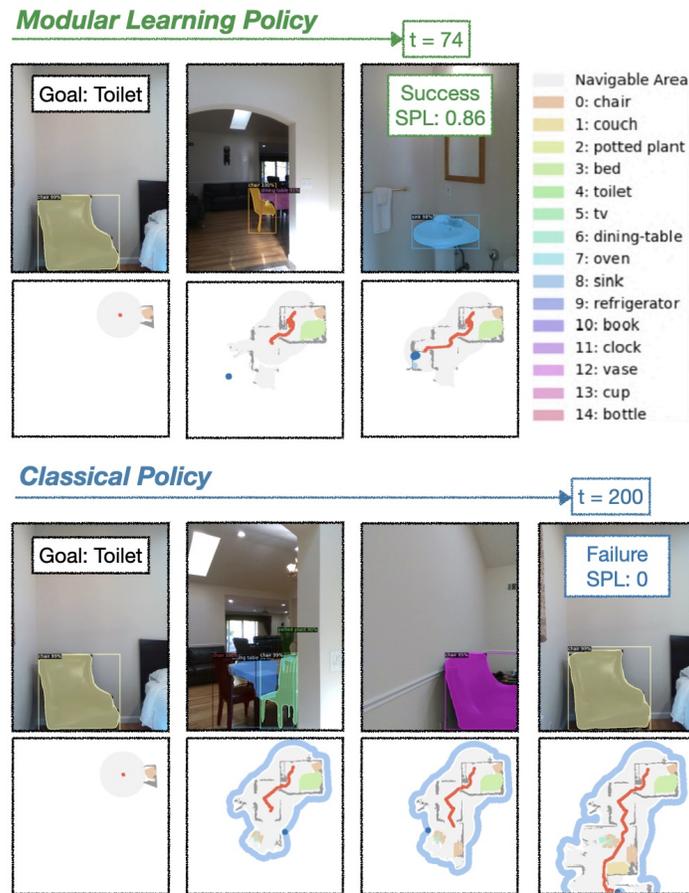


Fig. S1. Comparison of modular learning and classical policies. (A) Top: the learned semantic exploration of the modular learning approach finds the toilet goal in only 74 steps. (B) Bottom: the frontier exploration of the classical approach fails to find the toilet goal in 200 steps.

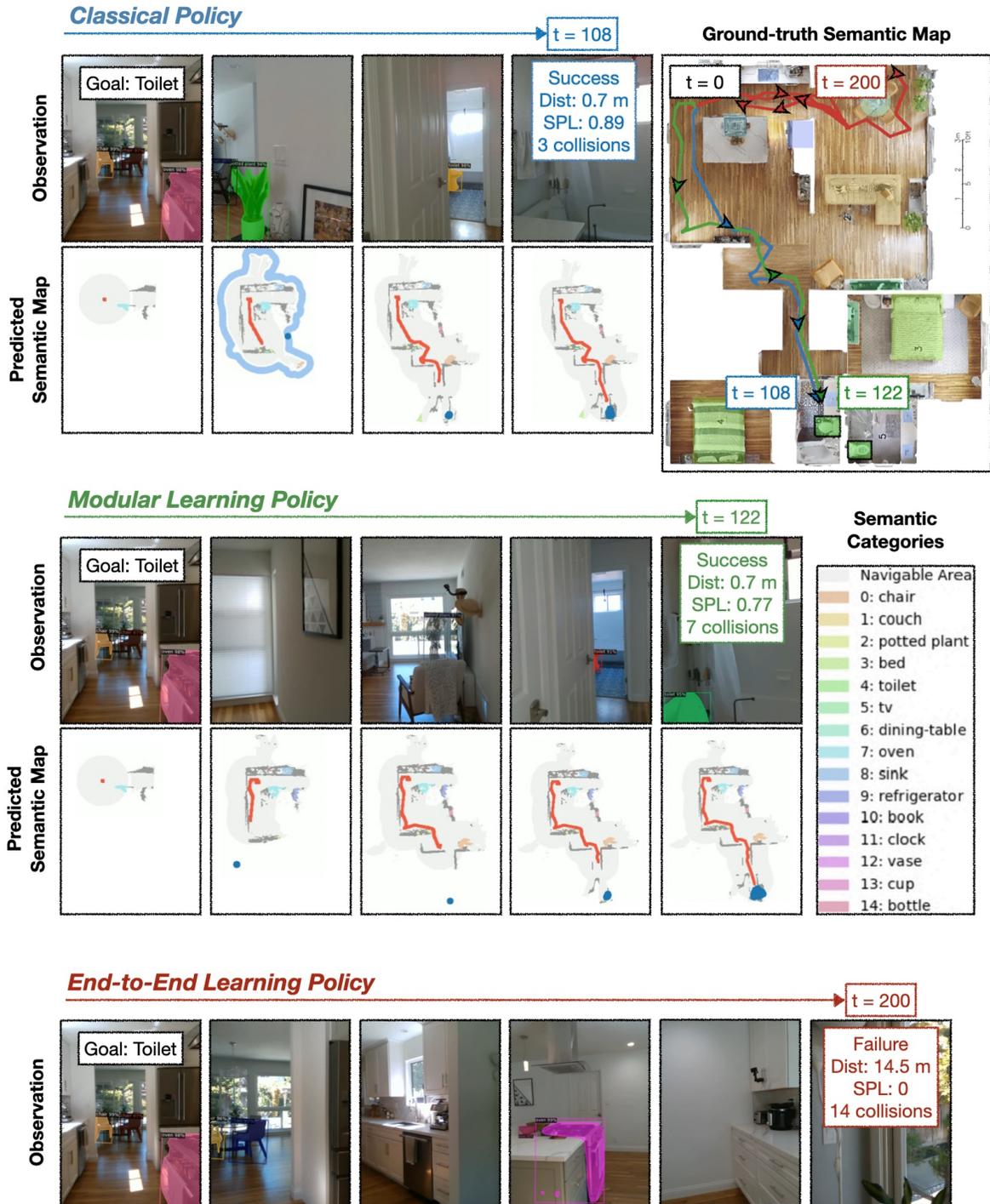


Fig. S2. Three approaches on the same episode. (A) Both the classical and modular learning policies reach the toilet goal efficiently, in 108 and 122 steps, respectively. **(B)** The end-to-end learning policy fails to explore beyond the kitchen and dining room and never reaches the toilet goal.



Fig. S3. Failure modes of end-to-end learning. (A) The end-to-end policy detects the toilet goal but fails to stop and heads back towards the living room. (B) The end-to-end policy fails to explore effectively, revisiting the same locations in the kitchen while looking for a TV.

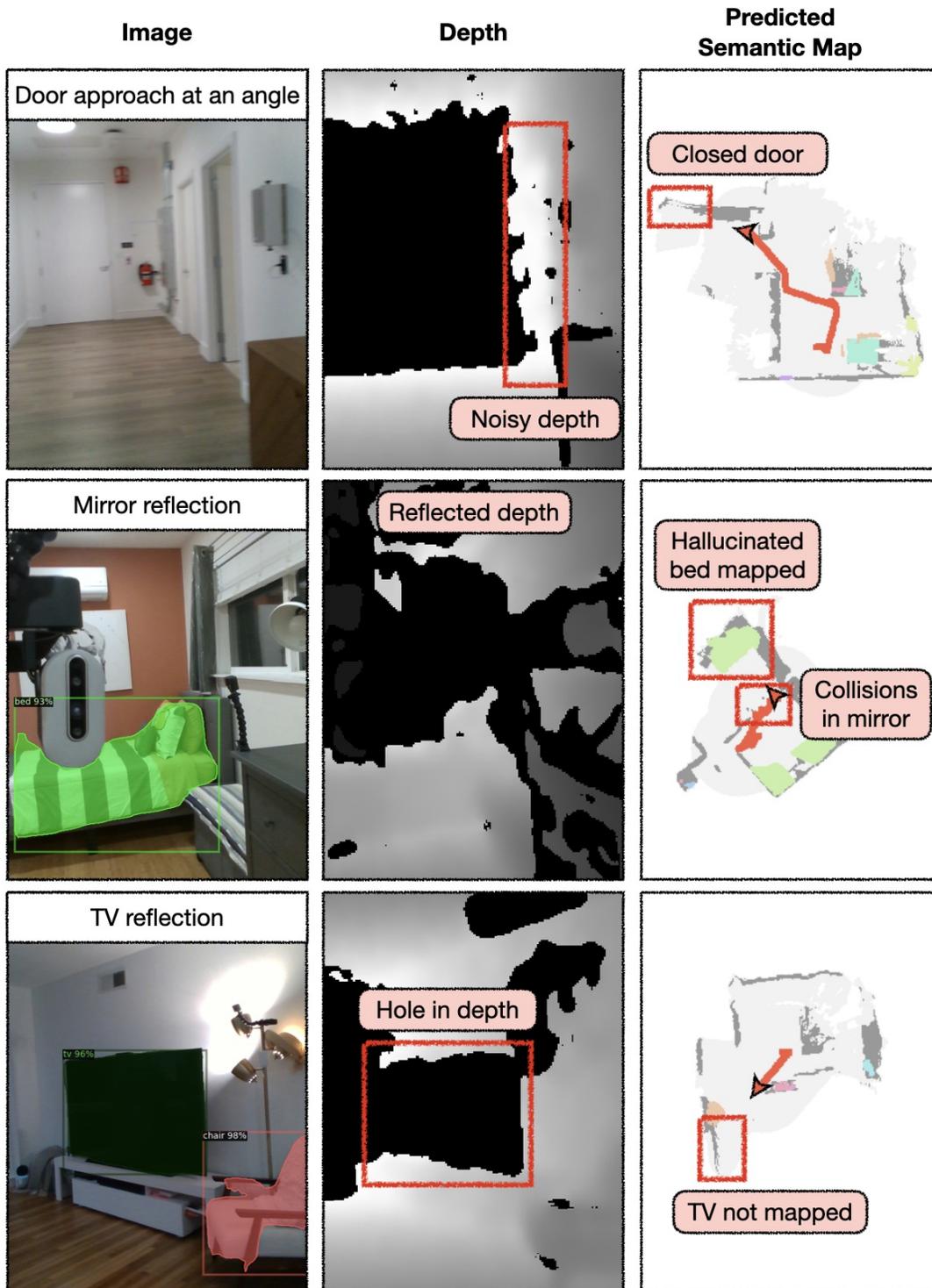
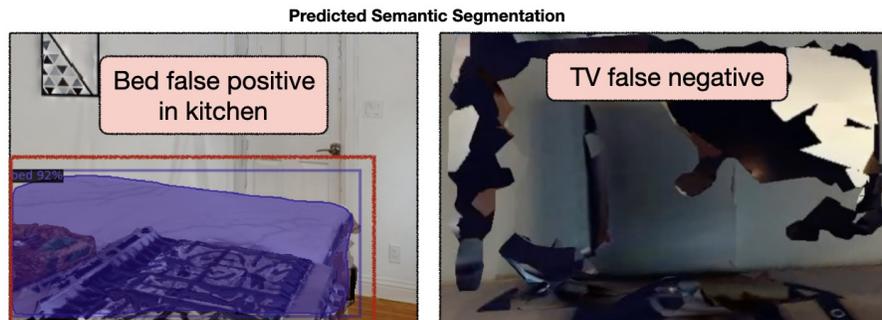


Fig. S4. Real-world depth sensor error modes. Top to bottom: (A) Depth noise walls a door when approaching at an angle. (B) Reflection in a mirror creates a duplicate bed on the map instead of an obstacle wall. (C) Reflection on a TV causes depth sensed beyond the sensor limit, and the TV is not mapped.

A - Visual Reconstruction Errors



B - Physical Reconstruction Errors

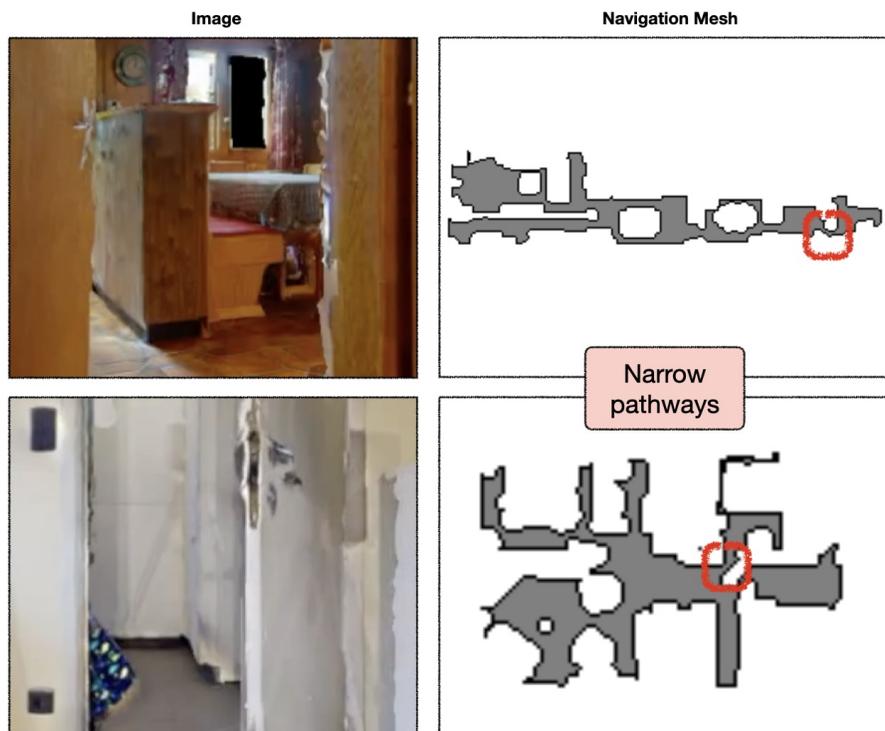


Fig. S5. Simulation visual and physical reconstruction error modes. (A) Visual reconstruction errors due to imperfect 3D scanning make segmentation errors more common in sim than reality. (B) Physical reconstruction errors lead to noisy navigation meshes with narrow pathways that are hard to navigate for discrete planners that work well in the real world.

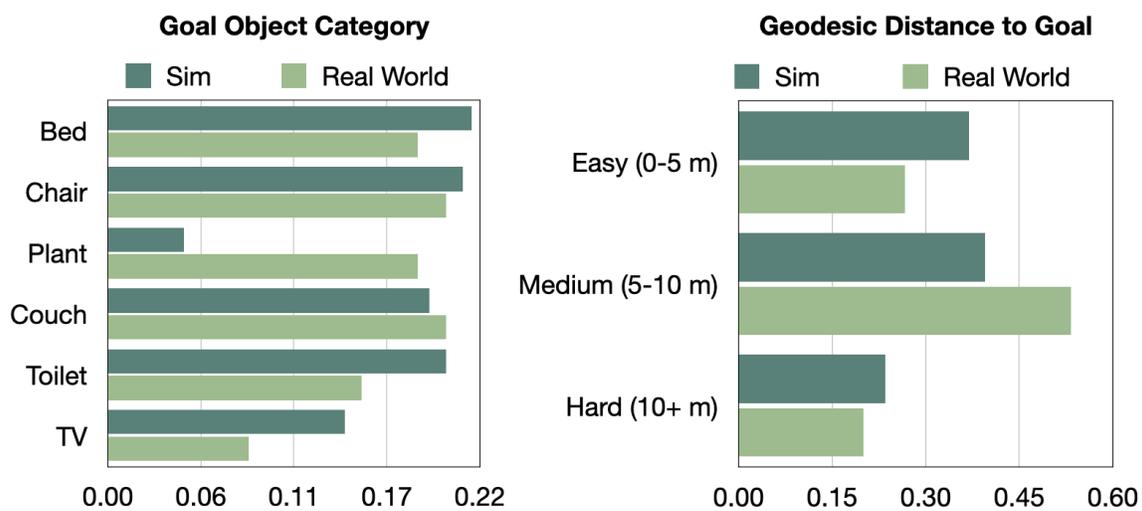


Fig. S6. Distributions of goal objects and geodesic distances to goal in reality vs. sim. (A) Distributions of goal objects and geodesic distances to the goal are comparable between a sim benchmark (val set of Habitat Challenge 2022) and our real-world evaluation setting.

Table S1. Navigation performance on all 60 episodes. (A) Details of navigation performance of classical, modular learning, and end-to-end learning approaches on 60 episodes across 6 object goal categories (“chair”, “couch”, “potted plant”, “toilet”, “tv”, and “bed”) in 6 homes.

Home	Episode		Classical		Modular Learning		End-to-end Learning	
	Goal Object	Shortest Path Length (m)	Success (SPL / Failure)	Collisions / Steps	Success (SPL / Failure)	Collisions / Steps	Success (SPL)	Collisions / Steps
1	couch2	3.9	✓ (0.64)	2 / 51	✓ (0.75)	1 / 39	✓ (0.76)	2 / 40
	chair2	4.8	✓ (0.98)	0 / 35	✓ (0.70)	0 / 45	✗	4 / 44
	plant2	6.3	✓ (0.77)	8 / 63	✓ (0.72)	1 / 64	✗	31 / 200
	chair1	6.7	✓ (0.48)	9 / 114	✗ (TV reflection)	25 / 145	✗	10 / 66
	tv1	7.5	✗ (TV reflection)	15 / 200	✓ (0.42)	8 / 125	✗	3 / 200
	couch1	8.1	✓ (0.33)	10 / 181	✓ (0.74)	2 / 78	✗	22 / 121
	plant1	9	✓ (0.60)	4 / 130	✓ (0.76)	6 / 127	✓ (0.99)	1 / 46
	toilet2	10.6	✓ (0.95)	3 / 72	✓ (0.66)	6 / 127	✗	16 / 200
	toilet1	13.3	✓ (0.89)	3 / 108	✓ (0.77)	7 / 122	✗	14 / 188
	bed1	14.2	✓ (0.80)	2 / 125	✓ (0.80)	7 / 135	✓ (0.51)	16 / 190
2	toilet2	3.5	✗ (mirror reflection)	37 / 126	✓ (0.49)	3 / 60	✗	18 / 180
	tv1	4.2	✗ (segmentation error)	7 / 136	✓ (0.99)	5 / 181	✗	4 / 203
	chair3	5.5	✓ (0.64)	1 / 66	✓ (0.97)	0 / 43	✓ (0.49)	5 / 68
	plant2	7.7	✓ (0.49)	3 / 125	✓ (0.63)	4 / 86	✗	45 / 138
	couch2	7.8	✓ (0.35)	8 / 156	✓ (0.66)	4 / 101	✗	11 / 200
	chair1	8.2	✓ (0.86)	3 / 79	✓ (0.79)	1 / 83	✗	11 / 200
	chair2	8.6	✓ (0.64)	12 / 103	✓ (0.47)	4 / 122	✗	4 / 104
	bed1	8.7	✓ (0.73)	7 / 75	✓ (0.67)	7 / 103	✗	22 / 189
	toilet1	9.1	✗ (exploration failure)	7 / 200	✗ (depth noise)	11 / 137	✗	1 / 98
	couch1	9.7	✓ (0.82)	2 / 90	✓ (0.74)	5 / 87	✗	30 / 200
plant1	9.8	✓ (0.51)	3 / 153	✓ (0.69)	2 / 94	✓ (0.51)	12 / 153	
3	plant2	3.1	✓ (0.76)	2 / 27	✓ (0.71)	1 / 28	✗	22 / 200
	tv1	4.2	✓ (0.71)	6 / 68	✓ (0.79)	6 / 41	✗	25 / 173
	bed1	4.5	✓ (0.72)	1 / 33	✓ (0.72)	1 / 33	✓ (0.27)	3 / 109
	bed2	5.2	✓ (0.49)	3 / 71	✓ (0.60)	3 / 69	✗	4 / 39
	chair1	5.9	✓ (0.99)	3 / 53	✓ (0.84)	2 / 54	✗	24 / 200
	couch1	9.3	✓ (0.69)	4 / 149	✓ (0.45)	8 / 160	✗	10 / 137
	plant1	10.2	✓ (0.46)	8 / 159	✓ (0.60)	4 / 121	✗	8 / 200
toilet1	10.3	✓ (0.83)	11 / 179	✗ (depth noise)	8 / 200	✗	23 / 200	
4	couch2	2	✓ (0.88)	0 / 19	✓ (0.99)	0 / 15	✗	32 / 94
	bed1	5.5	✗ (segmentation error)	6 / 133	✓ (0.39)	2 / 104	✓ (0.60)	2 / 68
	plant2	5.6	✗ (depth noise)	4 / 81	✓ (0.20)	10 / 187	✗	24 / 163
	couch1	6.1	✓ (0.90)	8 / 65	✓ (0.28)	6 / 154	✗	25 / 200
	bed2	6.4	✗ (segmentation error)	22 / 148	✓ (0.61)	3 / 82	✗	9 / 94
	chair1	7.2	✓ (0.86)	3 / 61	✓ (0.81)	1 / 63	✓ (0.31)	11 / 179
	chair2	7.3	✓ (0.82)	13 / 94	✓ (0.72)	6 / 69	✗	21 / 200
	plant1	8.3	✓ (0.70)	8 / 93	✓ (0.62)	8 / 104	✗	7 / 200
	tv1	9.7	✗ (TV reflection)	5 / 187	✗ (TV reflection)	30 / 188	✗	43 / 200
	plant3	10.2	✗ (depth noise)	15 / 163	✗ (depth noise)	6 / 101	✗	10 / 81
bed3	14.3	✓ (0.52)	4 / 152	✓ (0.90)	2 / 98	✓ (0.85)	12 / 101	
5	tv1	2	✓ (0.98)	1 / 17	✓ (0.89)	0 / 22	✓ (0.90)	1 / 19
	chair1	3.4	✓ (0.87)	1 / 27	✓ (0.97)	1 / 25	✓ (0.83)	8 / 34
	couch2	4.4	✓ (0.78)	0 / 36	✓ (0.82)	0 / 31	✗	9 / 45
	plant1	4.8	✓ (0.69)	2 / 59	✓ (0.96)	1 / 43	✗	16 / 79
	chair2	5	✓ (0.99)	0 / 33	✓ (0.99)	0 / 33	✗	14 / 98
	couch1	5.2	✓ (0.57)	4 / 71	✓ (0.77)	3 / 51	✗	19 / 134
	bed2	7.2	✓ (0.96)	3 / 58	✓ (0.92)	2 / 62	✓ (0.95)	10 / 60
	toilet1	8	✗ (exploration failure)	10 / 200	✓ (0.86)	3 / 89	✗	26 / 200
	bed1	8.7	✓ (0.82)	5 / 82	✓ (0.37)	6 / 163	✗	5 / 82
	toilet2	12.1	✗ (exploration failure)	10 / 200	✗ (exploration failure)	8 / 200	✗	17 / 200
6	toilet1	3.3	✓ (0.84)	2 / 29	✓ (0.90)	1 / 29	✓ (0.69)	6 / 39
	chair1	3.8	✓ (0.98)	1 / 32	✓ (0.92)	1 / 32	✓ (0.78)	7 / 45
	chair2	4.7	✓ (0.48)	3 / 64	✓ (0.94)	1 / 32	✗	12 / 87
	plant1	4.8	✓ (0.55)	2 / 60	✓ (0.36)	4 / 104	✗	9 / 70
	couch3	5.1	✓ (0.61)	5 / 76	✓ (0.61)	2 / 60	✗	18 / 149
	couch2	7.4	✓ (0.54)	4 / 112	✓ (0.79)	4 / 85	✗	16 / 122
	bed1	10.1	✓ (0.86)	1 / 94	✓ (0.83)	1 / 83	✗	33 / 200
	couch1	10.2	✓ (0.66)	1 / 125	✓ (0.53)	6 / 148	✗	24 / 200
	toilet2	10.9	✓ (0.58)	4 / 139	✓ (0.59)	6 / 187	✗	4 / 139
	bed2	13.1	✗ (depth noise)	4 / 200	✓ (0.45)	14 / 186	✗	26 / 200

Table S2. Navigation performance aggregated by home and goal object. We compare approaches by Success Rate (SR) and Success weighted by Path Length (SPL). **(A)** Results are consistent across homes. **(B)** Results are consistent across goal objects.

Home	Modular Learning SR (SPL)	Classical SR (SPL)	End-to-end Learning SR (SPL)
1	0.90 (0.63)	0.90 (0.64)	0.30 (0.23)
2	0.90 (0.61)	0.70 (0.44)	0.10 (0.05)
3	0.88 (0.59)	1.00 (0.70)	0.13 (0.03)
4	0.80 (0.49)	0.50 (0.39)	0.30 (0.17)
5	0.90 (0.76)	0.80 (0.67)	0.30 (0.27)
6	1.00 (0.69)	0.90 (0.61)	0.20 (0.15)

Goal	Modular Learning SR (SPL)	Classical SR (SPL)	End-to-end Learning SR (SPL)
Couch	1.00 (0.72)	1.00 (0.67)	0.08 (0.06)
Chair	0.90 (0.78)	1.00 (0.82)	0.33 (0.20)
Bed	1.00 (0.70)	0.70 (0.56)	0.45 (0.29)
Plant	0.90 (0.58)	0.80 (0.53)	0.18 (0.14)
Toilet	0.66 (0.47)	0.55 (0.45)	0.11 (0.08)
TV	0.83 (0.66)	0.50 (0.42)	0.20 (0.18)

Table S3. Modular learning real vs. sim error modes. (A) Real-world errors largely stem from depth sensor errors. (B) Sim errors from reconstruction errors: segmentation errors (more common in sim due to imperfect visual reconstruction) and navigation mesh errors (imperfect physical reconstruction).

Error mode	6 real homes	Sim replica	Sim benchmark
Segmentation error	0 / 6	2 / 2	10.1% / 18.6%
Navigation mesh error	-	0 / 2	5.5% / 18.6%
Exploration failure	1 / 6	0 / 2	3.0% / 18.6%
Depth sensor error:			
1 - Noise closes door	5 / 6	-	-
2 - Reflection (mirror, TV)			

Table S4. Modular learning programmatic error analysis on sim benchmark. We isolate (A) errors due to multi-floor navigation by comparing performance on all episodes and the subset of episodes with start and goal on the first floor, (B) errors due to segmentation failures by introducing ground-truth segmentation, (C) errors due to exploration failures by introducing a very large time budget (2000 steps).

Episode Set	Evaluation Condition	Successful Episodes
All episodes (2000)	500 steps budget, predicted segmentation	1078 / 2000
Episodes with start and goal on first floor (1093)	500 steps budget, predicted segmentation	733 / 1093
	500 steps budget, ground-truth segmentation	843 / 1093
	2000 steps budget, ground-truth segmentation	876 / 1093

Table S5. Modular learning manual analysis of remaining errors on sim benchmark. We manually classify the error mode for each of the remaining $1093 - 876 = 217$ episodes.

Error Mode	Episodes
Navmesh/planning error (agent stuck in narrow pathway)	60 / 217
Annotation error (agent reaches instance of goal category that is not annotated as belonging to the goal category)	157 / 217

Table S6. Modular learning episode outcomes on sim benchmark. We aggregate programmatic and manual error analyses among single-floor navigation episodes (1093 episodes).

Outcome	Episodes	Proportion
Segmentation error	110 (843 - 733)	10.1%
Navigation mesh error	60	5.5%
Exploration failure	33 (876 - 843)	3.0%
Success	890 (733 + 157)	81.4%

References

- [1] P. Anderson, *et al.*, On evaluation of embodied navigation agents, *arXiv:1807.06757* (2018).
- [2] H. P. Moravec, Obstacle avoidance and navigation in the real world by a seeing robot rover (Stanford University, 1980).
- [3] R. Chatila, J.-P. Laumond, Position referencing and consistent world modeling for mobile robots, *Proceedings. 1985 IEEE International Conference on Robotics and Automation* (IEEE, 1985), vol. 2, pp. 138–145.
- [4] A. Elfes, Sonar-based real-world mapping and navigation, *IEEE Journal on Robotics and Automation* **3**, 249 (1987).
- [5] A. Elfes, Using occupancy grids for mobile robot perception and navigation, *Computer* **22**, 46 (1989).
- [6] B. Kuipers, Y.-T. Byun, A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations, *Robotics and autonomous systems* **8**, 47 (1991).
- [7] I. J. Cox, Blanche—an experiment in guidance and navigation of an autonomous robot vehicle, *IEEE Transactions on robotics and automation* **7**, 193 (1991).
- [8] J. J. Leonard, H. F. Durrant-Whyte, I. J. Cox, Dynamic map building for an autonomous mobile robot, *The International Journal of Robotics Research* **11**, 286 (1992).
- [9] D. Fox, W. Burgard, S. Thrun, The dynamic window approach to collision avoidance, *IEEE Robotics & Automation Magazine* **4**, 23 (1997).
- [10] W. Burgard, *et al.*, Experiences with an interactive museum tour-guide robot, *Artificial intelligence* **114**, 3 (1999).
- [11] S. Thrun, *et al.*, MINERVA: A second-generation museum tour-guide robot, *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)* (IEEE, 1999), vol. 3.
- [12] S. Thrun, D. Fox, W. Burgard, F. Dellaert, Robust Monte Carlo localization for mobile robots, *Artificial intelligence* **128**, 99 (2001).
- [13] S. Thrun, *et al.*, Robotic mapping: A survey, *Exploring artificial intelligence in the new millennium* **1**, 1 (2002).
- [14] R. A. Newcombe, *et al.*, Kinectfusion: Real-time dense surface mapping and tracking, *2011 10th IEEE international symposium on mixed and augmented reality* (Ieee, 2011), pp. 127–136.

- [15] A. J. Davison, I. D. Reid, N. D. Molton, O. Stasse, MonoSLAM: Real-time single camera SLAM, *IEEE transactions on pattern analysis and machine intelligence* **29**, 1052 (2007).
- [16] E. S. Jones, S. Soatto, Visual-inertial navigation, mapping and localization: A scalable real-time causal approach, *The International Journal of Robotics Research* **30**, 407 (2011).
- [17] T. Sattler, *et al.*, Benchmarking 6dof outdoor visual localization in changing conditions, *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8601–8610.
- [18] B. Yamauchi, A frontier-based approach for autonomous exploration, *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)* (IEEE, 1997), pp. 146–151.
- [19] A. Flint, D. Murray, I. Reid, Manhattan scene understanding using monocular, stereo, and 3d features, *2011 International Conference on Computer Vision* (IEEE, 2011), pp. 2228–2235.
- [20] A. Kundu, Y. Li, F. Dellaert, F. Li, J. M. Rehg, Joint semantic segmentation and 3d reconstruction from monocular video, *European Conference on Computer Vision* (Springer, 2014), pp. 703–718.
- [21] S. L. Bowman, N. Atanasov, K. Daniilidis, G. J. Pappas, Probabilistic data association for semantic slam, *2017 IEEE international conference on robotics and automation (ICRA)* (IEEE, 2017), pp. 1722–1729.
- [22] L. Ma, J. Stückler, C. Kerl, D. Cremers, Multi-view deep learning for consistent semantic mapping with rgb-d cameras, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2017), pp. 598–605.
- [23] L. Zhang, *et al.*, Semantic SLAM based on object detection and improved octomap, *IEEE Access* **6**, 75545 (2018).
- [24] A. Rosinol, M. Abate, Y. Chang, L. Carlone, Kimera: an open-source library for real-time metric-semantic localization and mapping, *2020 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2020), pp. 1689–1696.
- [25] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, A. J. Davison, Slam++: Simultaneous localisation and mapping at the level of objects, *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013), pp. 1352–1359.
- [26] D. A. Pomerleau, Alvin: An autonomous land vehicle in a neural network, *Advances in neural information processing systems* **1** (1988).

- [27] U. Muller, J. Ben, E. Cosatto, B. Flepp, Y. Cun, Off-road obstacle avoidance through end-to-end learning, *Advances in neural information processing systems* **18** (2005).
- [28] V. Mnih, *et al.*, Human-level control through deep reinforcement learning, *Nature* **518**, 529 (2015).
- [29] T. P. Lillicrap, *et al.*, Continuous control with deep reinforcement learning, *arXiv preprint arXiv:1509.02971* (2015).
- [30] G. Lample, D. S. Chaplot, Playing FPS Games with Deep Reinforcement Learning, *The AAAI Conference on Artificial Intelligence (AAAI)* (2017).
- [31] Y. Zhu, *et al.*, Target-driven visual navigation in indoor scenes using deep reinforcement learning, *2017 IEEE international conference on robotics and automation (ICRA)* (IEEE, 2017), pp. 3357–3364.
- [32] P. Mirowski, *et al.*, Learning to navigate in complex environments, *International Conference on Learning Representations (ICLR)* (2017).
- [33] A. Dosovitskiy, V. Koltun, Learning to act by predicting the future, *International Conference on Learning Representations (ICLR)* (2017).
- [34] D. S. Chaplot, G. Lample, Arnold: An autonomous agent to play fps games, *The AAAI Conference on Artificial Intelligence (AAAI)* (2017).
- [35] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, V. Koltun, MINOS: Multimodal Indoor Simulator for Navigation in Complex Environments, *arXiv:1712.03931* (2017).
- [36] K. M. Hermann, *et al.*, Grounded language learning in a simulated 3D world, *arXiv:1706.06551* (2017).
- [37] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, R. Salakhutdinov, Gated-Attention Architectures for Task-Oriented Language Grounding, *arXiv:1706.07230* (2017).
- [38] P. Mirowski, *et al.*, Learning to navigate in cities without a map, *Advances in Neural Information Processing Systems* (2018), pp. 2419–2430.
- [39] F. Codevilla, M. Müller, A. López, V. Koltun, A. Dosovitskiy, End-to-end driving via conditional imitation learning, *2018 IEEE international conference on robotics and automation (ICRA)* (IEEE, 2018), pp. 4693–4700.
- [40] A. Banino, *et al.*, Vector-based navigation using grid-like representations in artificial agents, *Nature* **557**, 429 (2018).

- [41] J. Ye, D. Batra, A. Das, E. Wijmans, Auxiliary tasks and exploration enable objectgoal navigation, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16117–16126.
- [42] O. Maksymets, *et al.*, THDA: Treasure hunt data augmentation for semantic navigation, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 15374–15383.
- [43] R. Ramrakhya, E. Undersander, D. Batra, A. Das, Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 5173–5183.
- [44] E. Wijmans, *et al.*, Decentralized Distributed PPO: Solving PointGoal Navigation, *arXiv:1911.00357* (2019).
- [45] M. Deitke, *et al.*, ProcTHOR: Large-Scale Embodied AI Using Procedural Generation, *arXiv:2206.06994* (2022).
- [46] R. Brooks, A robust layered control system for a mobile robot, *IEEE journal on robotics and automation* **2**, 14 (1986).
- [47] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, J. Malik, Cognitive mapping and planning for visual navigation, *IEEE Computer Vision and Pattern Recognition Conference (CVPR)* (2017), pp. 2616–2625.
- [48] E. Parisotto, R. Salakhutdinov, Neural map: Structured memory for deep reinforcement learning, *International Conference on Learning Representations (ICLR)* (2018).
- [49] D. S. Chaplot, E. Parisotto, R. Salakhutdinov, Active neural localization, *International Conference on Learning Representations (ICLR)* (2018).
- [50] J. F. Henriques, A. Vedaldi, Mapnet: An allocentric spatial memory for mapping environments, *IEEE Computer Vision and Pattern Recognition Conference (CVPR)* (2018), pp. 8476–8484.
- [51] D. Gordon, *et al.*, Iqa: Visual question answering in interactive environments, *IEEE Computer Vision and Pattern Recognition Conference (CVPR)* (2018), pp. 4089–4098.
- [52] W. Yang, X. Wang, A. Farhadi, A. Gupta, R. Mottaghi, Visual semantic navigation using scene priors, *arXiv:1810.06543* (2018).
- [53] N. Savinov, A. Dosovitskiy, V. Koltun, Semi-parametric topological memory for navigation, *International Conference on Learning Representations* (2018).

- [54] N. Savinov, *et al.*, Episodic curiosity through reachability, *International Conference on Learning Representations (ICLR)* (2019).
- [55] R. McAllister, *et al.*, Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning (International Joint Conferences on Artificial Intelligence, Inc., 2017).
- [56] M. Müller, A. Dosovitskiy, B. Ghanem, V. Koltun, Driving policy transfer via modularity and abstraction, *arXiv:1804.09364* (2018).
- [57] D. Scaramuzza, *et al.*, Vision-controlled micro flying robots: from system design to autonomous navigation and mapping in GPS-denied environments, *IEEE Robotics & Automation Magazine* **21**, 26 (2014).
- [58] A. Mousavian, C. Eppner, D. Fox, 6-dof graspnet: Variational grasp generation for object manipulation, *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 2901–2910.
- [59] J. Mahler, *et al.*, Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, *arXiv preprint arXiv:1703.09312* (2017).
- [60] D. Morrison, *et al.*, Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge, *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 7757–7764.
- [61] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, R. Salakhutdinov, Learning To Explore Using Active Neural SLAM, *International Conference on Learning Representations (ICLR)* (2020).
- [62] D. S. Chaplot, D. P. Gandhi, A. Gupta, R. R. Salakhutdinov, Object goal navigation using goal-oriented semantic exploration, *Advances in Neural Information Processing Systems* **33**, 4247 (2020).
- [63] S. K. Ramakrishnan, Z. Al-Halah, K. Grauman, Occupancy anticipation for efficient exploration and navigation, *European Conference on Computer Vision* (Springer, 2020), pp. 400–418.
- [64] D. S. Chaplot, R. Salakhutdinov, A. Gupta, S. Gupta, Neural Topological SLAM for Visual Navigation, *IEEE Computer Vision and Pattern Recognition Conference (CVPR)* (2020).
- [65] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, K. Grauman, PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning, *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on* (IEEE, 2022).

- [66] M. Hahn, *et al.*, No RL, No Simulation: Learning to Navigate without Navigating, *Advances in Neural Information Processing Systems (NeurIPS)* (2021).
- [67] J. Krantz, A. Gokaslan, D. Batra, S. Lee, O. Maksymets, Waypoint Models for Instruction-Guided Navigation in Continuous Environments, *IEEE International Conference on Computer Vision (ICCV)* (2021).
- [68] D. An, *et al.*, 1st Place Solutions for RxR-Habitat Vision-and-Language Navigation Competition (CVPR 2022), *arXiv:2206.11610* (2022).
- [69] S. Y. Min, D. S. Chaplot, P. Ravikumar, Y. Bisk, R. Salakhutdinov, FILM: Following Instructions in Language with Modular Methods, *International Conference on Learning Representations (ICLR)* (2022).
- [70] X. Liu, H. Palacios, C. Muise, A Planning based Neural-Symbolic Approach for Embodied Instruction Following, *Interactions* **9**, 8 (2022).
- [71] M. Murray, M. Cakmak, Following Natural Language Instructions for Household Tasks with Landmark Guided Search and Reinforced Pose Adjustment, *IEEE Robotics and Automation Letters* (2022).
- [72] G. Sarch, *et al.*, TIDEE: Tidying Up Novel Rooms using Visuo-Semantic Commonsense Priors, *European Conference on Computer Vision (ECCV)* (2022).
- [73] B. Trabucco, G. Sigurdsson, R. Piramuthu, G. S. Sukhatme, R. Salakhutdinov, A Simple Approach for Visual Rearrangement: 3D Mapping and Semantic Search, *arXiv:2206.13396* (2022).
- [74] D. S. Chaplot, H. Jiang, S. Gupta, A. Gupta, Semantic Curiosity for Active Visual Learning, *ECCV* (2020).
- [75] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, R. Salakhutdinov, SEAL: Self-supervised Embodied Active Learning, *NeurIPS* (2021).
- [76] J. Duan, S. Yu, H. L. Tan, H. Zhu, C. Tan, A survey of embodied ai: From simulators to research tasks, *IEEE Transactions on Emerging Topics in Computational Intelligence* (2022).
- [77] D. Mishkin, A. Dosovitskiy, V. Koltun, Benchmarking Classic and Learned Navigation in Complex 3D Environments, *arXiv:1901.10915* (2019).
- [78] M. Savva, *et al.*, Habitat: A Platform for Embodied AI Research, *ICCV* (2019).
- [79] E. Kolve, *et al.*, AI2-THOR: An Interactive 3D Environment for Visual AI, *arXiv* (2022).

- [80] A. Kadian, *et al.*, Sim2real predictivity: Does evaluation in simulation predict real-world performance?, *IEEE Robotics and Automation Letters* **5**, 6670 (2020).
- [81] J. Truong, S. Chernova, D. Batra, Bi-directional domain adaptation for sim2real transfer of embodied navigation agents, *IEEE Robotics and Automation Letters* **6**, 2634 (2021).
- [82] J. Truong, *et al.*, Rethinking Sim2Real: Lower Fidelity Simulation Leads to Higher Sim2Real Transfer in Navigation, *arXiv preprint arXiv:2207.10821* (2022).
- [83] Z. Fu, *et al.*, Coupling vision and proprioception for navigation of legged robots, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17273–17283.
- [84] T. Miki, *et al.*, Learning robust perceptive locomotion for quadrupedal robots in the wild, *Science Robotics* **7**, eabk2822 (2022).
- [85] D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, S. Levine, Ving: Learning open-world navigation with visual goals, *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 13215–13222.
- [86] D. Shah, S. Levine, ViKiNG: Vision-Based Kilometer-Scale Navigation with Geographic Hints, *arXiv preprint arXiv:2202.11271* (2022).
- [87] P. Anderson, *et al.*, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 3674–3683.
- [88] S. Höfer, *et al.*, Sim2Real in robotics and automation: Applications and challenges, *IEEE transactions on automation science and engineering* **18**, 398 (2021).
- [89] A. Loquercio, *et al.*, Learning high-speed flight in the wild, *Science Robotics* **6**, eabg5810 (2021).
- [90] C. C. Kemp, A. Edsinger, H. M. Clever, B. Matulevich, The design of Stretch: A compact, lightweight mobile manipulator for indoor human environments, *IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2022), pp. 3150–3157.
- [91] D. Batra, *et al.*, Objectnav revisited: On evaluation of embodied agents navigating to objects, *arXiv:2006.13171* (2020).
- [92] K. Yadav, *et al.*, Habitat-Matterport 3D Semantics Dataset, *arXiv preprint arXiv:2210.05633* (2022).
- [93] A. Kadian, *et al.*, Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance?, *IEEE Robotics and Automation Letters (RA-L)* (2020).

- [94] S. K. Ramakrishnan, *et al.*, Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI, *arXiv:2109.08238* (2021).
- [95] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, I. Misra, Detecting twenty-thousand classes using image-level supervision, *arXiv preprint arXiv:2201.02605* (2022).
- [96] W. Li, X. Song, Y. Bai, S. Zhang, S. Jiang, ION: Instance-level Object Navigation, *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 4343–4352.
- [97] L. Pinto, A. Gupta, Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours, *2016 IEEE international conference on robotics and automation (ICRA)* (IEEE, 2016), pp. 3406–3413.
- [98] D. Kalashnikov, *et al.*, Scalable deep reinforcement learning for vision-based robotic manipulation, *Conference on Robot Learning* (PMLR, 2018), pp. 651–673.
- [99] J. Tobin, *et al.*, Domain randomization for transferring deep neural networks from simulation to the real world, *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on* (IEEE, 2017), pp. 23–30.
- [100] O. M. Andrychowicz, *et al.*, Learning dexterous in-hand manipulation, *The International Journal of Robotics Research* **39**, 3 (2020).
- [101] E. Kaufmann, *et al.*, Deep drone acrobatics, *arXiv preprint arXiv:2006.05768* (2020).
- [102] A. Szot, *et al.*, Habitat 2.0: Training home assistants to rearrange their habitat, *Advances in Neural Information Processing Systems (NeurIPS)* **34**, 251 (2021).
- [103] MetaAI, Fairo: A modular embodied agent architecture and platform for building embodied agents, <https://github.com/facebookresearch/fairo> (2021).
- [104] S. Kohlbrecher, J. Meyer, O. von Stryk, U. Klingauf, A Flexible and Scalable SLAM System with Full 3D Motion Estimation, *IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)* (IEEE, 2011).
- [105] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, *IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2980–2988.
- [106] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
- [107] F. Xia, *et al.*, Gibson env: Real-world perception for embodied agents, *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 9068–9079.

- [108] J. A. Sethian, A fast marching level set method for monotonically advancing fronts, *National Academy of Sciences* **93**, 1591 (1996).
- [109] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, *arXiv:1412.3555* (2014).
- [110] J. Jiang, L. Zheng, F. Luo, Z. Zhang, Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation, *arXiv:1806.01054* (2018).
- [111] S. Song, S. P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 567–576.
- [112] E. Wijmans, *et al.*, Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames, *arXiv:1911.00357* (2019).
- [113] S. Choi, Q.-Y. Zhou, V. Koltun, Robust reconstruction of indoor scenes, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 5556–5565.