

Cognitive Mapping and Planning for Visual Navigation

Saurabh Gupta · Varun Tolani · James Davidson
Sergey Levine · Rahul Sukthankar · Jitendra Malik

Received: date / Accepted: date

Abstract We introduce a neural architecture for navigation in novel environments. Our proposed architecture learns to map from first-person views and plans a sequence of actions towards goals in the environment. The Cognitive Mapper and Planner (CMP) is based on two key ideas: a) a unified joint architecture for mapping and planning, such that the mapping is driven by the needs of the task, and b) a spatial memory with the ability to plan given an incomplete set of observations about the world. CMP constructs a top-down belief map of the world and applies a differentiable neural net planner to produce the next action at each time step. The accumulated belief of the world enables the agent to track visited regions of the environment. We train and test CMP on navigation problems in simulation environments derived from scans of real world buildings. Our experiments demonstrate that CMP outperforms alternate learning-based architectures, as well as, classical mapping and path planning approaches in many cases. Furthermore, it naturally extends to semantically specified goals, such as “going to a chair”. We also deploy CMP on physical robots in indoor environments, where it achieves reasonable performance, even though it is trained entirely in simulation.

Keywords Visual Navigation · Spatial Representations · Learning for Navigation

Work done in part at Google, and in part at UC Berkeley. Project website with simulator, code, models, and videos: <https://sites.google.com/view/cognitive-mapping-and-planning/>.

S. Gupta^{1,2} V. Tolani¹ J. Davidson²

S. Levine^{1,2} R. Sukthankar¹ J. Malik^{1,2}

¹UC Berkeley, ²Google

{sgupta, slevine, malik}@eecs.berkeley.edu, vtolani@berkeley.edu, {jcdavidson, sukthankar}@google.com

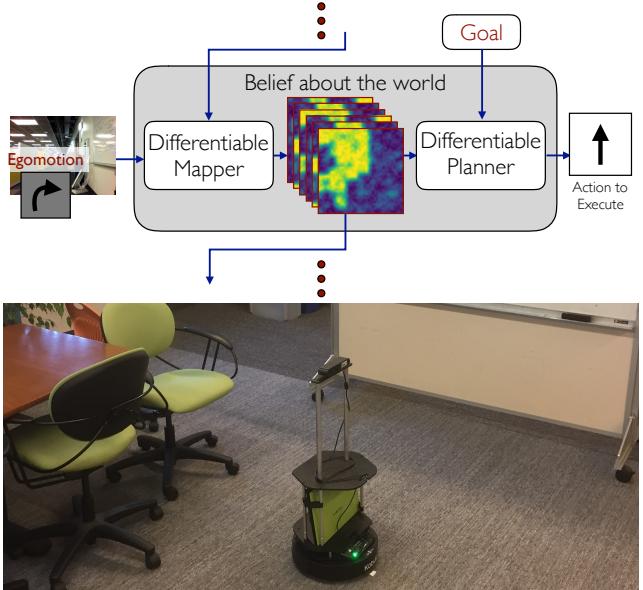


Fig. 1: Top: Network architecture: Our learned navigation network consists of mapping and planning modules. The mapper writes into a latent spatial memory that corresponds to an egocentric map of the environment, while the planner uses this memory alongside the goal to output navigational actions. The map is not supervised explicitly, but rather emerges naturally from the learning process. **Bottom:** We also describe experiments where we deploy our learned navigation policies on a physical robot.

1 Introduction

As humans, when we navigate through novel environments, we draw on our previous experience in similar conditions. We reason about free-space, obstacles and the topology of the environment, guided by common sense rules and heuristics for navigation. For example, to go from one room to another, I must first exit the initial room; to go to a room at the other end of the building, getting into a hallway is more

likely to succeed than entering a conference room; a kitchen is more likely to be situated in open areas of the building than in the middle of cubicles. The goal of this paper is to design a learning framework for acquiring such expertise, and demonstrate this for the problem of robot navigation in novel environments.

However, classic approaches to navigation rarely make use of such common sense patterns. Classical SLAM based approaches [19, 71] first build a 3D map using LIDAR, depth, or structure from motion, and then plan paths in this map. These maps are built purely geometrically, and nothing is known until it has been explicitly observed, even when there are obvious patterns. This becomes a problem for goal directed navigation. Humans can often guess, for example, where they will find a chair or that a hallway will probably lead to another hallway but a classical robot agent can at best only do uninformed exploration. The separation between mapping and planning also makes the overall system unnecessarily fragile. For example, the mapper might fail on texture-less regions in a corridor, leading to failure of the whole system, but precise geometry may not even be necessary if the robot just has to keep traveling straight.

Inspired by this reasoning, recently there has been an increasing interest in more end-to-end learning-based approaches that go directly from pixels to actions [54, 58, 80] without going through explicit model or state estimation steps. These methods thus enjoy the power of being able to learn behaviors from experience. However, it is necessary to carefully design architectures that can capture the structure of the task at hand. For instance Zhu *et al.* [80] use reactive memory-less vanilla feed forward architectures for solving visual navigation problems. In contrast, experiments by Tolman [72] have shown that even rats build sophisticated representations for space in the form of ‘cognitive maps’ as they navigate, giving them the ability to reason about shortcuts, something that a reactive agent is unable to.

This motivates our Cognitive Mapping and Planning (CMP) approach for visual navigation (Figure 1). CMP consists of a) a spatial memory to capture the layout of the world, and b) a planner that can plan paths given partial information. The mapper and the planner are put together into a unified architecture that can be trained to leverage regularities of the world. The mapper fuses information from input views as observed by the agent over time to produce a metric egocentric multi-scale belief about the world in a top-down view. The planner uses this multi-scale egocentric belief of the world to plan paths to the specified goal and outputs the optimal action to take. This process is repeated at each time step to convey the agent to the goal.

At each time step, the agent updates the belief of the world from the previous time step by a) using the ego-motion to transform the belief from the previous time step into the current coordinate frame and b) incorporating information

from the current view of the world to update the belief. This allows the agent to progressively improve its model of the world as it moves around. The most significant contrast with prior work is that our approach is trained end-to-end to take good actions in the world. To that end, instead of analytically computing the update to the belief (via classical structure from motion) we frame this as a learning problem and train a convolutional neural network to predict the update based on the observed first person view. We make the belief transformation and update operations differentiable thereby allowing for end-to-end training. This allows our method to adapt to the statistical patterns in real indoor scenes without the need for any explicit supervision of the mapping stage.

Our planner uses the metric belief of the world obtained through the mapping operation described above to plan paths to the goal. We use value iteration as our planning algorithm but crucially use a trainable, differentiable and hierarchical version of value iteration. This has three advantages, a) being trainable it naturally deals with partially observed environments by explicitly learning when and where to explore, b) being differentiable it enables us to train the mapper for navigation, and c) being hierarchical it allows us to plan paths to distant goal locations in time complexity that is logarithmic in the number of steps to the goal.

Our approach is a reminiscent of classical work in navigation that also involves building maps and then planning paths in these maps to reach desired target locations. However, our approach differs from classical work in the following significant way: except for the architectural choice of maintaining a metric belief, everything else is learned from data. This leads to some very desirable properties: a) our model can learn statistical regularities of indoor environments in a task-driven manner, b) jointly training the mapper and the planner makes our planner more robust to errors of the mapper, and c) our model can be used in an online manner in novel environments without requiring a pre-constructed map.

This paper originally appeared at CVPR 2017. In this journal article, we additionally describe real world deployment of our learned policies on a TurtleBot 2 platform, and report results of our deployment on indoor test environments. We have also incorporated feedback from the community. In particular, we have added comparisons to a policy that very closely resembles a classical mapping and planning method. We have also included more visualizations of the representations produced by the mapper. Finally, we situate the current work in context of the new research directions that are emerging in the field at the intersection of machine learning, robotics and computer vision.

2 Related Work

Navigation is one of the most fundamental problems in mobile robotics. The standard approach is to decompose the

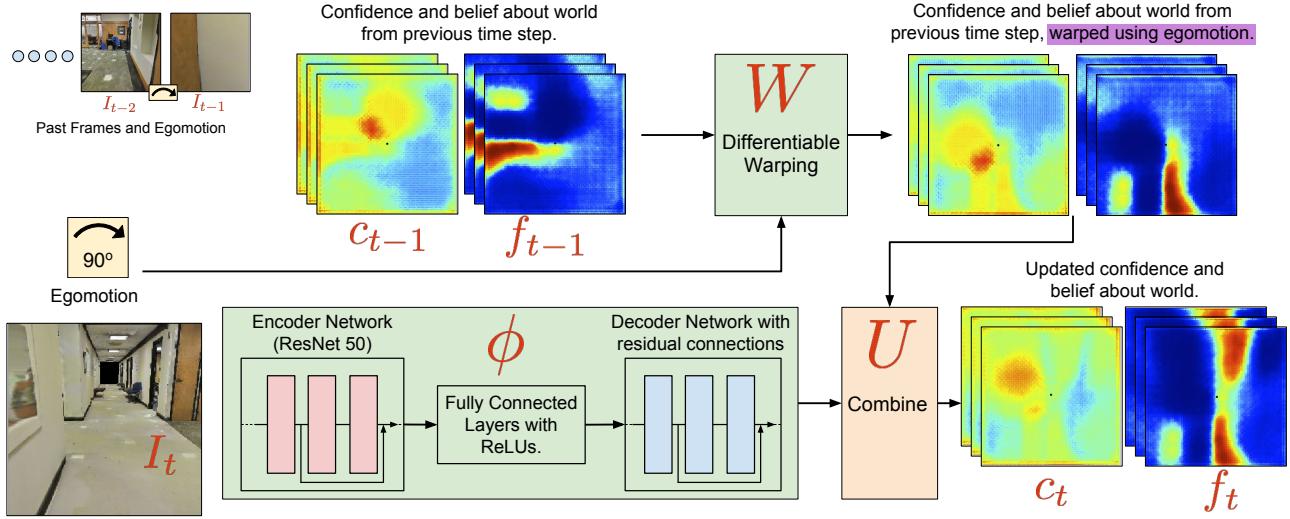


Fig. 2: Architecture of the mapper: The mapper module processes first person images from the robot and integrates the observations into a latent memory, which corresponds to an **egocentric map of the top-view** of the environment. The mapping operation is not supervised explicitly – the mapper is free to write into memory whatever information is most useful for the planner. In addition to filling in obstacles, the mapper also stores confidence values in the map, which allows it to make probabilistic predictions about unobserved parts of the map by exploiting learned patterns.

problem into two separate stages: (1) mapping the environment, and (2) planning a path through the constructed map [22, 46]. Decomposing navigation in this manner allows each stage to be developed independently, **but prevents each from exploiting the specific needs of the other**. A comprehensive survey of classical approaches for mapping and planning can be found in [71].

Mapping has been well studied in computer vision and robotics in the form of structure from motion and simultaneous localization and mapping [25, 37, 40, 69] with a variety of sensing modalities such as range sensors, RGB cameras and RGB-D cameras. These approaches take a purely geometric approach. Learning based approaches [32, 78] study the problem in isolation thus only learning generic task-independent maps. Path planning in these inferred maps has also been well studied, with pioneering works from Canny [13], Kavraki *et al.* [44] and LaValle and Kuffner [53]. Works such as [23, 24] have studied the joint problem of mapping and planning. While this relaxes the need for pre-mapping by incrementally updating the map while navigating, but still treat navigation as a purely geometric problem, Konolige *et al.* [50] and Aydemir *et al.* [8] proposed approaches which leveraged semantics for more informed navigation. Kuipers *et al.* [52] introduce a cognitive mapping model using hierarchical abstractions of maps. Semantics have also been associated with 3D environments more generally [29, 51].

As an alternative to separating out discrete mapping and planning phases, reinforcement learning (RL) methods directly learn policies for robotic tasks [47, 49, 62]. A major challenge with using RL for this task is the need to process complex sensory input, such as camera images. Recent works in deep reinforcement learning (DRL) learn policies in an end-to-end manner [58] going from pixels to actions.

Follow-up works [28, 57, 68] propose improvements to DRL algorithms, [35, 57, 60, 74, 79] study how to incorporate memory into such neural network based models. We build on the work from Tamar *et al.* [70] who study how explicit planning can be incorporated in such agents, but do not consider the case of first-person visual navigation, nor provide a framework for memory or mapping. [60] study the generalization behavior of these algorithms to novel environments they have not been trained on.

In context of navigation, learning and DRL has been used to obtain policies [3, 16, 26, 43, 60, 70, 73, 80]. Some of these works [26, 43], focus on the problem of learning controllers for effectively maneuvering around obstacles directly from raw sensor data. Others, such as [11, 60, 70], focus on the planning problem associated with navigation under full state information [70], designing strategies for faster learning via episodic control [11], or incorporate memory into DRL algorithms to ease generalization to new environments. Most of this research (except [80]) focuses on navigation in synthetic mazes which have little structure to them. Given these environments are randomly generated, the policy learns a random exploration strategy, but has no statistical regularities in the layout that it can exploit. We instead test on layouts obtained from real buildings, and show that our architecture consistently outperforms feed forward and LSTM models used in prior work.

The research most directly relevant to our work is the contemporary work of Zhu *et al.* [80]. Similar to us, Zhu *et al.* also study first-person view navigation using macro-actions in more realistic environments instead of synthetic mazes. Zhu *et al.* propose a feed forward model which when trained in one environment can be finetuned in another environment. Such a memory-less agent cannot map, plan or explore the

environment, which our expressive model naturally does. Zhu *et al.* also don't consider zero-shot generalization to previously unseen environments, and focus on smaller worlds where memorization of landmarks is feasible. In contrast, we explicitly handle generalization to new, never before seen interiors, and show that our model generalizes successfully to floor plans not seen during training.

Relationship to contemporary research. In this paper, we used scans of real world environments to construct visually realistic simulation environments to study representations that can enable navigation in novel previously unseen environments. Since conducting this research, over the last year, there has been a major thrust in this direction in computer vision and related communities. Numerous works such as [4, 14, 17] have collected large-scale datasets consisting of scans of real world environments, while [66, 75, 76] have built more sophisticated simulation environments based on such scans. A related and parallel stream of research studies whether or not models trained in simulators can be effectively transferred to the real world [12, 64], and how the domain gap between simulation and the real world may be reduced [76]. A number of works have studied related navigation problems in such simulation environments [6, 15, 36]. Researchers have also gone beyond specifying goals as a desired location in space and finding objects of interest as done in this paper, for example, Wu *et al.* [75] generalize the goal specification to also include rooms of interest, and Das *et al.* [18] allow goal specification via templated questions. Finally, a number of works have also pursued the problem of building representation for space in context of navigation. [10, 27, 45, 61] use similar 2D spatial representations, Mirowski *et al.* [56] use fully-connected LSTMs, while Savinov *et al.* [65] develop topological representations. Interesting reinforcement learning techniques have also been explored for the task of navigation [21, 56].

3 Problem Setup

To be able to focus on the high-level mapping and planning problem we remove confounding factors arising from low-level control by conducting our experiments in simulated real world indoor environments. Studying the problem in simulation makes it easier to run exhaustive evaluation experiments, while the use of scanned real world environments allows us to retain the richness and complexity of real scenes. We also only study the static version of the problem, though extensions to dynamic environments would be interesting to explore in future work.

We model the robot as a cylinder of a fixed radius and height, equipped with vision sensors (RGB cameras or depth cameras) mounted at a fixed height and oriented at a fixed pitch. The robot is equipped with low-level controllers which

provide relatively high-level macro-actions $\mathcal{A}_{x,\theta}$. These macro-actions are a) stay in place, b) rotate left by θ , c) rotate right by θ , and d) move forward x cm, denoted by a_0, a_1, a_2 and a_3 , respectively. We further assume that the environment is a grid world and the robot uses its macro-actions to move between nodes on this graph. The robot also has access to its precise egomotion. This amounts to assuming perfect visual odometry [59], which can itself be learned [31], but we defer the joint learning problem to future work.

We want to learn policies for this robot for navigating in *novel* environments that it has not previously encountered. We study two navigation tasks, a *geometric task* where the robot is required to go to a target location specified in robot's coordinate frame (*e.g.* 250cm forward, 300cm left) and a *semantic task* where the robot is required to go to an object of interest (*e.g.* a chair). These tasks are to be performed in novel environments, neither the exact environment map nor its topology is available to the robot.

Our navigation problem is defined as follows. At a given time step t , let us assume the robot is at a global position (position in the world coordinate frame) P_t . At each time step the robot receives as input the image of the environment \mathcal{E} , $I_t = I(\mathcal{E}, P_t)$ and a target location $(x_t^g, y_t^g, \theta_t^g)$ (or a semantic goal) specified in the coordinate frame of the robot. The navigation problem is to learn a policy that at every time steps uses these inputs (current image, egomotion and target specification) to output the action that will convey the robot to the target as quickly as possible.

Experimental Testbed. We conduct our experiments on the Stanford large-scale 3D Indoor Spaces (S3DIS) dataset introduced by Armeni *et al.* [7]. The dataset consists of 3D scans (in the form of textured meshes) collected in 6 large-scale indoor areas that originate from 3 different buildings of educational and office use. The dataset was collected using the Matterport scanner [1]. Scans from 2 buildings were used for training and the agents were tested on scans from the 3rd building. We pre-processed the meshes to compute space traversable by the robot. We also precompute a directed graph $\mathcal{G}_{x,\theta}$ consisting of the set of locations the robot can visit as nodes and a connectivity structure based on the set of actions $\mathcal{A}_{x,\theta}$ available to the robot to efficiently generate training problems. More details in Section A4.

4 Mapping

We describe how the mapping portion of our learned network can integrate first-person camera images into a top-down 2D representation of the environment, while learning to leverage statistical structure in the world. Note that, unlike analytic mapping systems, the map in our model amounts to a latent representation. Since it is fed directly into a learned planning module, it need not encode purely free space representations, but can instead function as a general spatial memory. The

model learns to store inside the map whatever information is most useful for generating successful plans. However to make description in this section concrete, we assume that the mapper predicts free space.

The mapper architecture is illustrated in Figure 2. At every time step t we maintain a cumulative estimate of the free space f_t in the coordinate frame of the robot. f_t is represented as a multi-channel 2D feature map that metrically represents space in the top-down view of the world. f_t is estimated from the current image I_t , cumulative estimate from the previous time step f_{t-1} and egomotion between the last and this step e_t using the following update rule:

$$f_t = U(W(f_{t-1}, e_t), f'_t) \quad \text{where, } f'_t = \phi(I_t). \quad (1)$$

here, W is a function that transforms the free space prediction from the previous time step f_{t-1} according to the egomotion in the last step e_t , ϕ is a function that takes as input the current image I_t and outputs an estimate of the free space based on the view of the environment from the current location (denoted by f'_t). U is a function which accumulates the free space prediction from the current view with the accumulated prediction from previous time steps. Next, we describe how each of the functions W , ϕ and U are realized.

The function W is realized using bi-linear sampling. Given the ego-motion, we compute a backward flow field $\rho(e_t)$. This backward flow maps each pixel in the current free space image f_t to the location in the previous free space image f_{t-1} where it should come from. This backward flow ρ can be analytically computed from the ego-motion (as shown in Section A1). The function W uses bi-linear sampling to apply this flow field to the free space estimate from the previous frame. Bi-linear sampling allows us to back-propagate gradients from f_t to f_{t-1} [42], which will make it possible to train this model end to end.

The function ϕ is realized by a convolutional neural network. Because of our choice to represent free space always in the coordinate frame of the robot, this becomes a relatively easy function to learn, given the network only has to output free space in the current coordinate, rather than in an arbitrary world coordinate frame determined by the cumulative egomotion of the robot so far.

Intuitively, the network can use semantic cues (such as presence of scene surfaces like floor and walls, common furniture objects like chairs and tables) alongside other learned priors about size and shapes of common objects to generate free space estimates, even for object that may only be partially visible. Qualitative results in Section A2 show an example for this where our proposed mapper is able to make predictions for spaces that haven't been observed.

The architecture of the neural network that realizes function ϕ is shown in Figure 2. It is composed of a convolutional encoder which uses residual connections [33] and produces a representation of the scene in the 2D image space. This

representation is transformed into one that is in the egocentric 2D top-down view via fully connected layers. This representation is up-sampled using up-convolutional layers (also with residual connections) to obtain the update to the belief about the world from the current frame.

In addition to producing an estimate of the free space from the current view f'_t the model also produces a confidence c'_t . This estimate is also warped by the warping function W and accumulated over time into c_t . This estimate allows us to simplify the update function, and can be thought of as playing the role of the update gate in a gated recurrent unit. The update function U takes in the tuples (f_{t-1}, c_{t-1}) , and (f'_t, c'_t) and produces (f_t, c_t) as follows:

$$f_t = \frac{f_{t-1}c_{t-1} + f'_t c'_t}{c_{t-1} + c'_t} \quad \text{and} \quad c_t = c_{t-1} + c'_t \quad (2)$$

We chose an analytic update function to keep the overall architecture simple. This can be replaced with more expressive functions like those realized by LSTMs [38].

Mapper performance in isolation. To demonstrate that our proposed mapper architecture works we test it in isolation on the task of free space prediction. Section A2 shows qualitative and quantitative results.

5 Planning

Our planner is based on value iteration networks proposed by Tamar *et al.* [70], who observed that a particular type of planning algorithm called value iteration [9] can be implemented as a neural network with alternating convolutions and channel-wise max pooling operations, allowing the planner to be differentiated with respect to its inputs. Value iteration can be thought of as a generalization of Dijkstra's algorithm, where the value of each state is iteratively recalculated at each iteration by taking a max over the values of its neighbors plus the reward of the transition to those neighboring states. This plays nicely with 2D grid world navigation problems, where these operations can be implemented with small 3×3 kernels followed by max-pooling over channels. Tamar *et al.* [70] also showed that this reformulation of value iteration can also be used to learn the planner (the parameters in the convolutional layer of the planner) by providing supervision for the optimal action for each state. Thus planning can be done in a trainable and differentiable manner by very deep convolutional network (with channel wise max-pooling). For our problem, the mapper produces the 2D top-view of the world which shares the same 2D grid world structure as described above, and we use value iteration networks as a trainable and differentiable planner.

Hierarchical Planning. Value iteration networks as presented in [70](v2) are impractical to use for any long-horizon planning problem. This is because the planning step size is coupled with the action step size thus leading to a) high

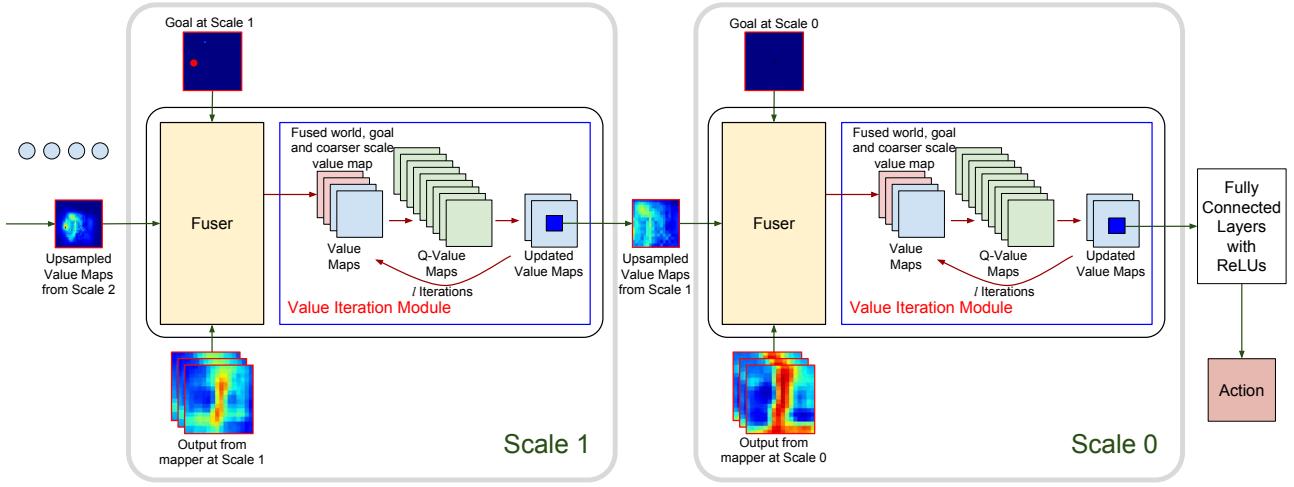


Fig. 3: Architecture of the hierarchical planner: The hierarchical planner takes the egocentric multi-scale belief of the world output by the mapper and uses value iteration expressed as convolutions and channel-wise max-pooling to output a policy. The planner is trainable and differentiable and back-propagates gradients to the mapper. The planner operates at multiple scales (scale 0 is the finest scale) of the problem which leads to efficiency in planning.

computational complexity at run time, and b) a hard learning problem as gradients have to flow back for as many steps. To alleviate this problem, we extend the [hierarchical version](#) presented in [70](v1).

Our hierarchical planner [plans at multiple spatial scales](#). We start with a k times spatially downsampled environment and conduct l value iterations in this downsampled environment. The output of this value iteration process is center cropped, upsampled, and used for doing value iterations at a finer scale. This process is repeated to finally reach the resolution of the original problem. This procedure allows us to plan for goals which are as far as $l2^k$ steps away while performing (and backpropagating through) only lk planning iterations. This efficiency increase comes [at the cost of approximate planning](#).

Planning in Partially Observed Environments. Value iteration networks have only been evaluated when the environment is fully observed, i.e. the entire map is known while planning. However, for our navigation problem, the map is only partially observed. Because the planner is not hand specified but learned from data, it can learn policies which naturally take partially observed maps into account. Note that the mapper produces not just a belief about the world but also an uncertainty c_i , the planner knows which parts of the map have and haven't been observed.

6 Joint Architecture

Our final architecture, Cognitive Mapping and Planning (CMP) puts together the mapper and planner described above. At each time step, the mapper updates its multi-scale belief about the world based on the current observation. This up-

dated belief is input to the planner [which outputs the action to take](#). As described previously, all parts of the network are differentiable and allow for end-to-end training, and [no additional direct supervision is used to train the mapping module](#) – rather than producing maps that match some ground truth free space, the mapper produces maps that allow the planner to choose effective actions.

Training Procedure. We optimize the CMP network with [fully supervised training using DAGGER](#) [63]. We generate training trajectories by sampling arbitrary start and goal locations on the graph $\mathcal{G}_{x,\theta}$. We generate supervision for training by computing shortest paths on the graph. We use an online version of DAGGER, where during each episode we sample the next state based on the action from the agent's current policy, or from the expert policy. We use scheduled sampling and anneal the probability of sampling from the expert policy using inverse sigmoid decay.

Note that [the focus of this work is on studying different architectures](#) for navigation. Our proposed architecture can also be trained with alternate paradigms for learning such policies, such as reinforcement learning. We chose DAGGER for training our models because we found it to be significantly [more sample efficient and stable in our domain, allowing us to focus on the architecture design](#).

7 Experiments

The goal of this paper is to learn policies for visual navigation for different navigation tasks in novel indoor environments. We first describe these different navigation tasks, and performance metrics. We then discuss different comparison points that quantify the novelty of our test environments, difficulty

of tasks at hand. Next, we compare our proposed CMP architecture to other learning-based methods and to classical mapping and planning based methods. We report all numbers on the test set. The test set consists of a floor from an altogether different building not contained in the training set. (See dataset website and Section A4 for environment visualizations.)

Tasks. We study two tasks: a *geometric task*, where the goal is to reach a point in space, and a *semantic task*, where the goal is to find objects of interest. We provide more details about both these tasks below:

1. **Geometric Task:** The goal is specified geometrically in terms of position of the goal in robot’s coordinate frame. Problems for this task are generated by sampling a start node on the graph and then sampling an end node which is within 32 steps from the starting node and preferably in another room or in the hallway (we use room and hallway annotations from the dataset [7]). This is same as the *PointGoal* task as described in [5].
2. **Semantic Task:** We consider three tasks: ‘go to a chair’, ‘go to a door’ and ‘go to a table’. The agent receives a one-hot vector indicating the object category it must go to and is considered successful if it can reach *any* instance of the indicated object category. We use object annotations from the S3DIS dataset [7] to setup this task. We initialize the agent such that it is within 32 time steps of at least one instance of the indicated category, and train it to go towards the nearest instance. This is same as the *ObjectGoal* task as described in [5].

The same sampling process is used during training and testing. For testing, we sample 4000 problems on the test set. The test set consists of a floor from an altogether different building not contained in the training set. These problems remain fixed across different algorithms that we compare. We measure performance by measuring the distance to goal after running the policies for a maximum number of time steps (200), or if they emit the *stop* action.

Performance Metrics. We report multiple performance metrics: a) the mean distance to goal, b) the 75th percentile distance to goal, and c) the success rate (the agent succeeds if it is within a distance of 3 steps of the goal location) as a function of number of time-steps. We plot these metrics as a function of time-steps and also report performance at 39 and 199 time steps in the various tables. For the most competitive methods, we also report the SPL metric¹ (higher is better)

¹ For computing SPL, we use the shortest-path on the graph as the shortest-path length. We count both rotation and translation actions for both the agent’s path and the shortest path. An episode is considered successful, if the agent ends up within 3 steps of the goal location. For the geometric task, we run the agent till it outputs the ‘stay-in-place’ action, or for a maximum of 200 time steps. For the semantic task, we train a separate ‘stop’ predictor. This ‘stop’ predictor is trained to predict if the agent is within 3 steps of the goal or not. The probability at which the episode should be terminated is determined on the validation set.

as introduced in [5]. In addition to measuring whether the agent reaches the goal, SPL additionally also measures the efficiency of the path used and whether the agent reliably determines that it has reached the goal or not.

Training Details. Models are trained asynchronously using TensorFlow [2]. We used ADAM [48] to optimize our loss function and trained for 60K iterations with a learning rate of 0.001 which was dropped by a factor of 10 every 20K iterations (we found this necessary for consistent training across different runs). We use weight decay of 0.0001 to regularize the network and use batch-norm [39]. We use ResNet-50 [34] pre-trained on ImageNet [20] to represent RGB images. We transfer supervision from RGB images to depth images using cross modal distillation [30] between RGB-D image pairs rendered from meshes in the training set to obtain a pre-trained ResNet-50 model to represent depth images.

7.1 Baselines

Our experiments are designed to test performance at visual navigation in novel environments. We first quantify the differences between training and test environments using a nearest neighbor trajectory method. We next quantify the difficulty of our environments and evaluation episodes by training a blind policy that only receives the relative goal location at each time step. Next, we test the effectiveness of our memory-based architecture. We compare to a purely reactive agent to understand the role of memory for this task, and to a LSTM-based policy to test the effectiveness of our specific memory architecture. Finally, we make comparisons with classical mapping and planning based techniques. Since the goal of this paper is to study various architectures for navigation we train all these architectures the same way using DAGGER [63] as described earlier. We provide more details for each of these baselines below.

1. **Nearest Neighbor Trajectory Transfer:** To quantify similarity between training and testing environments, we transfer optimal trajectories from the train set to the test set using visual nearest neighbors (in RGB ResNet-50 feature space). At each time step, we pick the location in the training set which results in the most similar view to that seen by the agent at the current time step. We then compute the optimal action that conveys the robot to the same relative offset in the training environment from this location and execute this action at the current time step. This procedure is repeated at each time step. Such a transfer leads to very poor results.
2. **No image, goal location only with LSTM:** Here, we ignore the image and simply use the relative goal location (in robot’s current coordinate frame) as input to a LSTM,

and predict the action that the agent should take. The relative goal location is embedded into a K dimensional space via fully connected layers with ReLU non-linearities before being input to the LSTM.

3. **Reactive Policy, Single Frame:** We next compare to a reactive agent that uses the first-person view of the world. As described above we use ResNet-50 to extract features. These features are passed through a few fully connected layers, and combined with the representation for the relative goal location which is used to predict the final action. We experimented with additive and multiplicative combination strategies and both performed similarly.
4. **Reactive Policy, Multiple Frames:** We also consider the case where the reactive policy receives 3 previous frames in addition to the current view. Given the robot’s step-size is fairly large we consider a late fusion architecture and fuse the information extracted from ResNet-50. Note that this architecture is similar to the one used in [80]. The primary differences are: goal is specified in terms of relative offset (instead of an image), training uses DAGGER (which utilizes denser supervision) instead of A3C, and testing is done in novel environments. These adaptations are necessary to make an interpretable comparison on our task.
5. **LSTM Based Agent:** Finally, we also compare to an agent which uses an LSTM based memory. We introduce LSTM units on the multiplicatively combined image and relative goal location representation. Such an architecture also gives the LSTM access to the egomotion of the agent (via how the relative goal location changes between consecutive steps). Thus this model has access to all the information that our method uses. We also experimented with other LSTM based models (ones without egomotion, inputting the egomotion more explicitly, etc.), but weren’t able to reliably train them in early experiments and did not pursue them further.
6. **Purely Geometric Mapping:** We also compare to a purely geometric incremental mapping and path planning policy. We projected observed 3D points incrementally into a top-down occupancy map using the ground truth egomotion and camera extrinsics and intrinsics. When using depth images as input, these 3D points are directly available. When using RGB images as input, we triangulated SIFT feature points in the RGB images (registered using the ground truth egomotion) to obtain the observed 3D points (we used the COLMAP library [67]). This occupancy map was used to compute a grid-graph (unoccupied cells are assumed free). For the geometric task, we mark the goal location on this grid-graph and execute the action that minimizes the distance to the goal node. For the semantic task, we use purely geometric exploration

along with a semantic segmentation network trained² to identify object categories of interest. The agent systematically explores the environment using frontier-based exploration [77]³ till it detects the specified category using the semantic segmentation network. These labels are projected onto the occupancy map using the 3D points, and nodes in the grid-graph are labelled as goals. We then output the action that minimizes the distance to these inferred goal nodes.

For this baseline, we experimented with different input image sizes, and increased the frequency at which RGB or depth images were captured. We validated a number of other hyper-parameters: a) number of points in a cell before it is considered occupied, b) number of intervening cell to be occupied before it is considered non-traversable, c) radius for morphological opening of the semantic labels on the map. 3D reconstruction from RGB images was computationally expensive, and thus we report comparisons to these classical baselines on a subset of the test cases.

7.2 Results

Geometric Task. We first present results for the geometric task. Figure 4 plots the error metrics over time (for 199 time steps), while Table 1 reports these metrics at 39 and 199 time steps, and SPL (with a max episode length of 199). We summarize the results below:

1. We first note that nearest neighbor trajectory transfer does not work well, with the mean and median distance to goal being 22 and 25 steps respectively. This highlights the differences between the train and test environments in our experiments.
2. Next, we note that the ‘No Image LSTM’ baseline performs poorly as well, with a success rate of 6.2% only. This suggests that our testing episodes aren’t trivial. They don’t just involve going straight to the goal, but require understanding the layout of the given environment.
3. Next, we observe that the reactive baseline with a single frame also performs poorly, succeeding only 8.2% of the time. Note that this reactive baseline is able to perform well on the training environments obtaining a mean distance to goal of about 9 steps, but perform poorly on the test set only being able to get to within 17 steps of the goal on average. This suggests that a reactive agent is able to effectively memorize the environments it was trained on, but fails to generalize to novel environments, this is not

² We train this semantic segmentation network to segment chairs, doors and table on the S3DIS dataset [7].

³ We sample a *goal* location outside the map, and try to reach it, as for the geometric task. As there is no path to this location, the agent ends up systematically exploring the environment.

Method	Mean		75 th %ile		Success %age		Mean		75 th %ile		Success %age		SPL %age		
	RGB	Depth	RGB	Depth	RGB	Depth	RGB	Depth	RGB	Depth	RGB	Depth	RGB	Depth	
Geometric Task (4000 episodes)															
Initial	25.3	25.3	30	30	0.7	0.7									
No Image LSTM	20.8	20.8	28	28	6.2	6.2									
Reactive (1 frame)	20.9	17.0	28	26	8.2	21.9									
Reactive (4 frames)	17.0	8.9	26	18	20.3	56.1	16.7	8.2	26	17	22.8	62.2	17.4	52.0	
LSTM	10.3	5.9	21	6	52.1	71.3	8.1	3.5	18	0	69.5	88.5	51.3	69.1	
Our (CMP)	7.7	4.8	14	1	63.0	78.8	5.4	3.3	0	0	80.0	89.3	59.4	73.7	
Geometric Task (1000 episodes)															
Classical (900px, 4× images)	20.3	3.3	29	2	17.4	89.6	20.4	3.3	29	2	17.7	90.7	15.9	80.6	
Our (CMP)	7.7	5.2	14	1	63.0	78.4	5.3	3.5	0	0	80.6	88.7	59.6	73.1	
Our (CMP [+6 MP3D Envs])	6.3	2.8	7	0	71.5	86.1	3.8	1.9	0	0	89.7	94.6	70.8	82.3	

Table 1: Results for the Geometric Task: We report the mean distance to goal location, 75th percentile distance to goal, success rate and SPL for various methods for the geometric task. Unshaded table reports metrics at time step 39, while the shaded table reports metrics for selected competitive methods at time step 199. Top part of the table reports comparisons between learning-based methods, while the bottom part report comparisons to classical purely geometry-based methods.

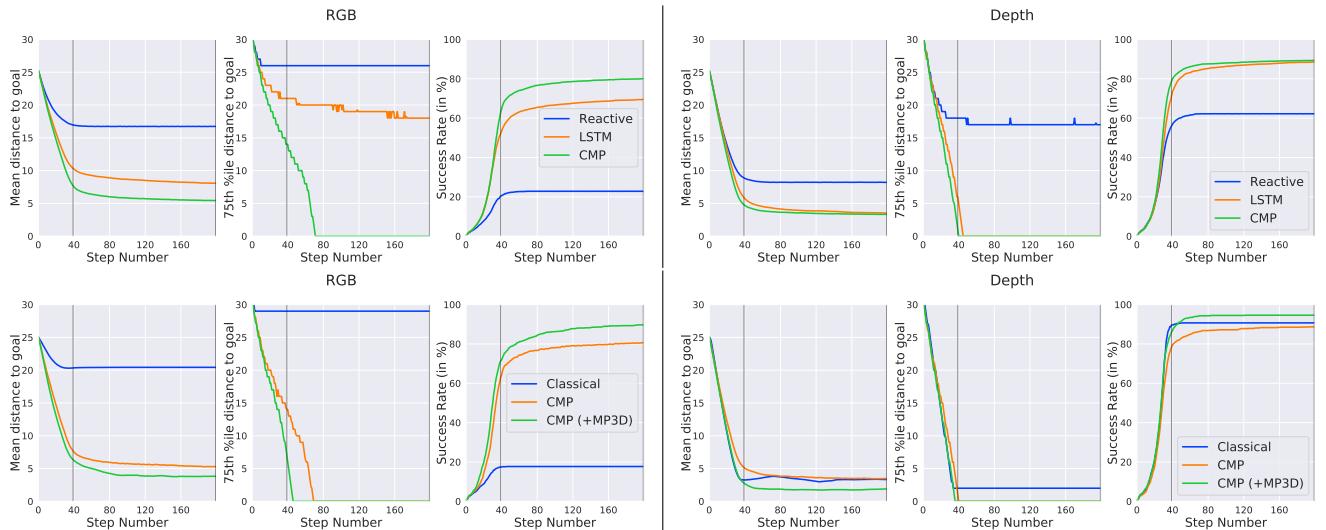


Fig. 4: Geometric Task: We plot the mean distance to goal, 75th percentile distance to goal (lower is better) and success rate (higher is better) as a function of the number of steps. Top row compares the 4 frame reactive agent, LSTM based agent and our proposed CMP based agent when using RGB images as input (left three plots) and when using depth images as input (right three plots). Bottom row compares classical mapping and planning with CMP (again, left is with RGB input and right with depth input). We note that CMP outperforms all these baselines, and using depth input leads to better performance than using RGB input.

- surprising given it does not have any form of memory to allow it to map or plan. We also experimented with using Drop Out in the fully connected layers for this model but found that to hurt performance on both the train and the test sets.
- Using additional frames as input to the reactive policy leads to a large improvement in performance, and boosts performance to 20%, and to 56% when using depth images.
 - The LSTM based model is able to consistently outperform these reactive baseline across all metrics. This indicates that memory does have a role to play in navigation in novel environments.

- Our proposed method **CMP**, outperforms all of these learning based methods, across all metrics and input modalities. CMP achieves a lower 75th %ile distance to goal (14 and 1 as compared to 21 and 5 for the LSTM) and improves the success rate to 62.5% and 78.3% from 53.0% and 71.8%. CMP also obtains higher SPL (59.4% vs. 51.3% and 73.7% vs. 69.1% for RGB and depth input respectively).
- We next compare to classical mapping and path planning. We first note that a purely geometric approach when provided with depth images does really well, obtaining a SPL of 80.6%. Access to depth images and perfect pose allows efficient and accurate mapping, leading to high performance. In contrast, when using only RGB images

Method	Success %age [39 steps]				Success %age [199 steps]				SPL %age [199 steps]			
	Chair	Door	Table	Mean	Chair	Door	Table	Mean	Chair	Door	Table	Mean
RGB (4000 episodes)												
Initial	10.7	12.1	10.3	11.1	10.7	12.1	10.3	11.1				
Reactive (4 frames)	24.8	24.7	18.6	22.7	19.1	24.9	19.7	21.2	12.7	13.3	10.2	12.1
LSTM	22.7	30.9	19.1	24.2	24.7	32.0	21.5	26.0	7.8	14.6	6.9	9.7
Our (CMP)	25.0	40.2	25.7	30.3	46.9	44.6	30.1	40.5	11.9	18.9	10.0	13.6
RGB (500 episodes)												
Classical (Explore + Sem. Segm.)	11.6	28.3	13.6	17.8	16.2	29.1	22.2	22.5	2.5	10.1	5.1	5.9
Our (CMP)	17.9	41.5	25.9	28.4	35.3	44.6	32.1	37.3	9.5	18.8	13.6	14.0
<i>Our (CMP [+ 6 MP3D Envs])</i>	42.8	40.3	21.0	34.7	80.3	42.2	40.7	54.4	21.1	20.7	15.1	19.0
Depth (4000 episodes)												
Reactive (4 frames)	28.4	23.6	29.2	27.1	33.8	23.8	38.7	32.1	14.0	9.8	11.9	11.9
LSTM	22.1	28.9	26.8	25.9	26.0	30.6	31.2	29.3	10.6	15.1	13.2	13.0
Our (CMP)	48.4	41.3	34.6	41.5	66.6	42.6	43.6	51.0	22.1	20.0	17.4	19.8
Depth (1000 episodes)												
Classical (Explore + Sem. Segm.)	38.0	72.4	28.5	46.3	38.9	49.8	43.0	43.9	27.5	24.7	24.3	25.5
Our (CMP)	50.2	41.9	34.6	42.2	67.7	43.4	44.7	51.9	22.4	19.9	17.9	20.1
<i>Our (CMP [+ 6 MP3D Envs])</i>	43.9	43.2	42.5	43.2	60.4	64.5	66.5	63.8	19.5	22.9	26.1	22.8

Table 2: Results for Semantic Task: We report success rate at 39 and 199 time steps, and SPL. We report performance for individual categories as well as their average. Top part reports comparisons with RGB input, bottom part reports comparisons with depth input. We compare learning based methods and a classical baseline (based on exploration and semantic segmentation). We also report performance of CMP when trained with more data (+6 MP3D Envs).

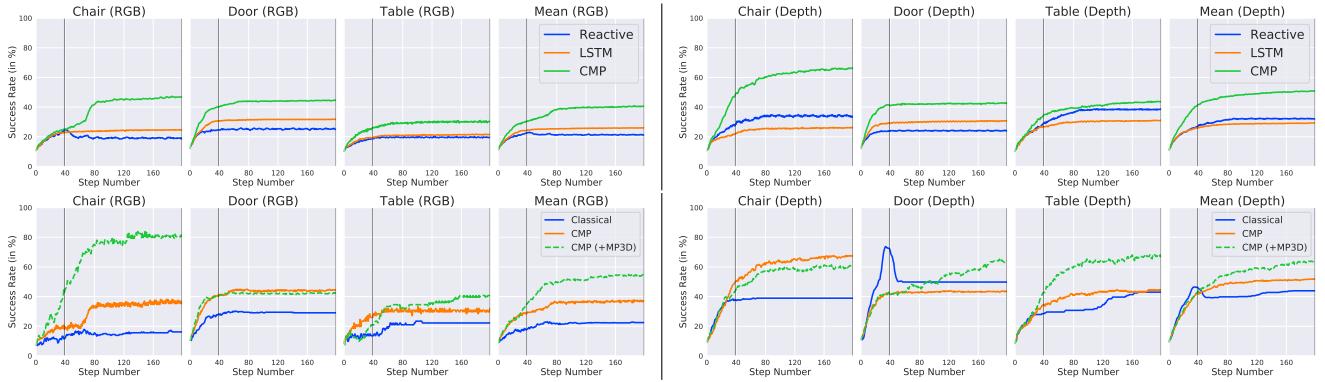


Fig. 5: Semantic Task: We plot the success rate as a function of the number of steps for different categories. Top row compares learning based approaches (4 frame reactive agent, LSTM based agent and our proposed CMP based agent). Bottom row compares a classical approach (using exploration along with semantic segmentation) and CMP. Left plots show performance when using RGB input, right plots show performance with depth input. See text for more details.

as input (but still with perfect pose), performance drops sharply to only 15.9%. There are two failure modes: spurious stray points in reconstruction that get treated as obstacles, and failure to reconstruct texture-less obstacles (such as walls) and bumping into them. In comparison, CMP performs well even when presented with just RGB images, at 59.6% SPL. Furthermore, when CMP is trained with more data (6 additional large buildings from the Matterport3D dataset [14]), performance improves further, to 70.8% SPL for RGB input and 82.3% SPL for depth input. Though we tried our best at implementing the classical purely geometry-based method, we note that they may be improved further by introducing and validating over more

and more hyper-parameters, specially for the case where depth observations are available as input.

Variance Over Multiple Runs. We also report variance in performance over five re-trainings from different random initializations of the network for the 3 most competitive methods (Reactive with 4 frames, LSTM, and CMP) for the depth image case. Figure A4 shows the performance, the solid line shows the median metric value and the surrounding shaded region represents the minimum and maximum metric value over the five re-trainings. As we are using imitation learning (instead of reinforcement learning) for training our models, variation in performance is reasonably small for all models and CMP leads to significant improvements.

Ablations. We also studied ablated versions of our proposed method. We summarize the key takeaways, a learned mapper leads to better navigation performance than an analytic mapper, planning is crucial (specially for when using RGB images as input) and single-scale planning works slightly better than the multi-scale planning at the cost of increased planning cost. More details in Section A3.

Additional comparisons between LSTM and CMP. We also conducted additional experiments to further compare the performance of the LSTM baseline with our model in the most competitive scenario where both methods use depth images. We summarize the key conclusions here and provide more details in Section A3. We report performance when the target is much further away (64 time steps away) in Table A3 (top). We see a larger gap in performance between LSTM and CMP for this test scenarios. We also compared performance of CMP and LSTM over problems of different difficulty and observed that CMP is generally better across all values of hardness, but for RGB images it is particularly better for cases with high hardness (Figure A3). We also evaluate how well these models generalize when trained on a single scene, and when transferring across datasets. We find that there is a smaller drop in performance for CMP as compared to LSTM (Table A3 (bottom)). More details in Section A3. Figure 6 visualizes and discusses some representative success and failure cases for CMP, video examples are available on the project website.

Semantic Task. We next present results for the semantic task, where the goal is to find object of interest. The agent receives a one-hot vector indicating the object category it must go to and is considered successful if it can reach any instance of the indicated object category. We compare our method to the best performing reactive and LSTM based baseline models from the geometric navigation task⁴. This is a challenging task specially because the agent may start in a location from which the desired object is not visible, and it must learn to explore the environment to find the desired object. Figure 5 and Table 2 reports the success rate and the SPL metric for the different categories we study. Figure 6 shows sample trajectories for this task for CMP. We summarize our findings below:

1. This is a hard task, performance for all methods is much lower than for the geometric task of reaching a specified point in space.
2. CMP performs better than the other two learning based baselines across all metrics.
3. Comparisons to the classical baseline of geometric exploration followed by use of semantic segmentation (Fig-

⁴ This LSTM is impoverished because it no longer receives the egomotion of the agent as input (because the goal can not be specified as an offset relative to the robot). We did experiment with a LSTM model which received egomotion as input but weren't able to train it in initial experiments.

ure 5 (bottom) orange vs. blue line) are also largely favorable to CMP. Performance for classical baseline with RGB input suffers due to inaccuracy in estimating the occupancy of the environment. With depth input, this becomes substantially easier, leading to better performance. A particularly interesting case is that of finding doors. As the classical baseline explores, it comes close to doors as it exits the room it started from. However, it is unable to stop (possibly being unable to reliably detect them). This explains the spike in performance in Figure 5.

4. We also report SPL for this task for the different methods in Table 2. We observe that though the success rates are high, SPL numbers are low. In comparison to success rate, SPL additionally measures path efficiency and whether the agent is able to reliably determine that it has reached the desired target. Figure 5 (bottom) shows that the success rate continues to improve over the length of the episodes, implying that the agent does realize that it has reached the desired object of interest. Thus, we believe SPL numbers are low because of inefficiency in reaching the target, specially as SPL measures efficiency with respect to the closest object of interest using full environment information. This can be particularly strict in novel environments where the agent may not have any desired objects in view, and thus needs to explore the environment to be able to find them. Nevertheless, CMP outperforms learning-based methods on this metric, and also outperforms our classical baseline when using RGB input.
5. CMP when trained with additional data (6 additional buildings from the Matterport3D dataset [14]) performs much better (green vs. orange lines in Figure 5 (bottom)), indicating scope for further improvements in such policies as larger datasets become available. Semantic segmentation networks for the classical baseline can similarly be improved using more data (possibly also from large-scale Internet datasets), but we leave those experiments and comparisons for future work.

7.3 Visualizations

We visualize activations at different layers in the CMP network to check if the architecture conforms to the intuitions that inspired the design of the network. We check for the following three aspects: a) is the representation produced by the mapper indeed spatial, b) does the mapper capture anything beyond what a purely geometric mapping pipeline captures, and c) do the value maps obtained from the value iteration module capture the behaviour exhibited by the agent.

Is the representation produced by the mapper spatial? We train simple readout functions on the learned mapper representation to predict free space around the agent. Figure 7 visualizes the output of these readout functions at

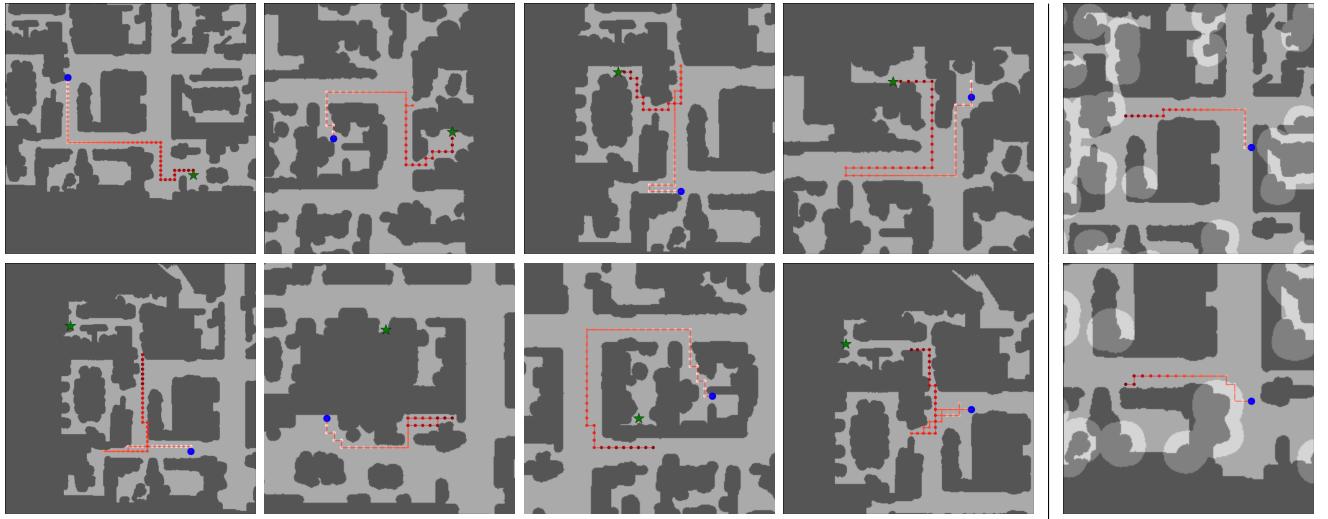


Fig. 6: Representative Success and Failure Cases for CMP: We visualize trajectories for some typical success and failure cases for CMP. Dark gray regions show occupied space, light gray regions show free space. The agent starts from the blue dot and is required to reach the green star (or semantic regions shown in light gray). The agent’s trajectory is shown by the dotted red line. While we visualize the trajectories in the top view, note that the agent only receives the first person view as input. **Top plots** show success cases for geometric task. We see that the agent is able to traverse large distances across multiple rooms to get to the target location, go around obstacles and quickly resolve that it needs to head to the next room and not the current room. The last two plots show cases where the agent successfully backtracks. **Bottom plots** show failure cases for geometric task: problems with navigating around tight spaces (entering through a partially opened door, and getting stuck in the corner (the gap is not big enough to pass through)), missing openings which would have lead to shorter paths, thrashing around in space without making progress. **Right plots** visualize trajectories for ‘go to the chair’ semantic task. The top figure shows a success case, while the bottom figure shows a typical failure case where the agent walks right through a chair region.

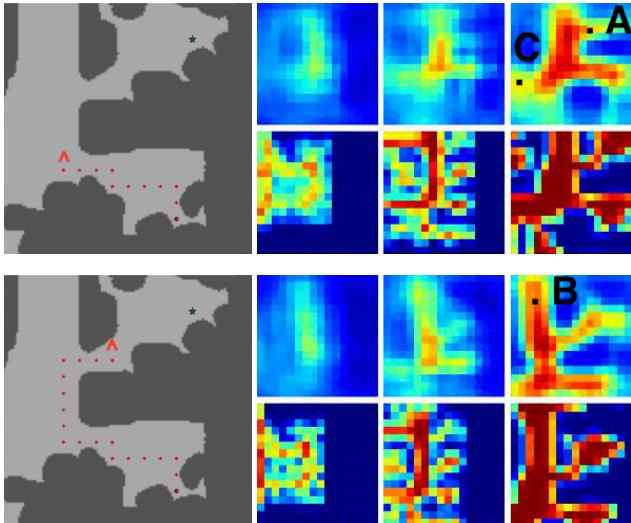


Fig. 7: We visualize the output of the map readout function trained on the representation learned by the mapper (see text for details) as the agent moves around. The two rows show two different time steps from an episode. For each row, the gray map shows the current position and orientation of the agent (red \wedge), and the locations that the agent has already visited during this episode (red dots). The top three heatmaps show the output of the map readout function and the bottom three heatmaps show the ground truth free space at the three scales used by CMP (going from coarse to fine from left to right). We observe that the readout maps capture the free space in the regions visited by the agent (room entrance at point A, corridors at points B and C).

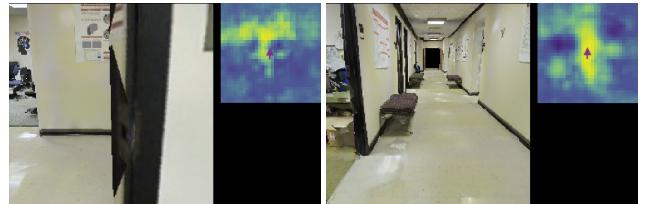


Fig. 8: We visualize first-person images and the output of the readout function output for free-space prediction derived from the representation produced by the mapper module (in egocentric frame, that is the agent is at the center looking upwards (denoted by the purple arrow)). In the left example, we can make a prediction behind the wall, and in the right example, we can make predictions inside the room.

two time steps from an episode as the agent moves. We see that the representation produced by the mapper is in correspondence with the actual free space around the agent. The representation produced by the mapper is indeed spatial in nature. We also note that readouts are generally better at finer scales.

What does the mapper representation capture? We next try to understand as to what information is captured in these spatial representations. First, as discussed above the representation produced by the mapper can be used to predict free space around the agent. Note that the agent was never trained to predict free space, yet the representations produced by the mapper carry enough information to predict free space reasonable well. Second, Figure 8 shows free

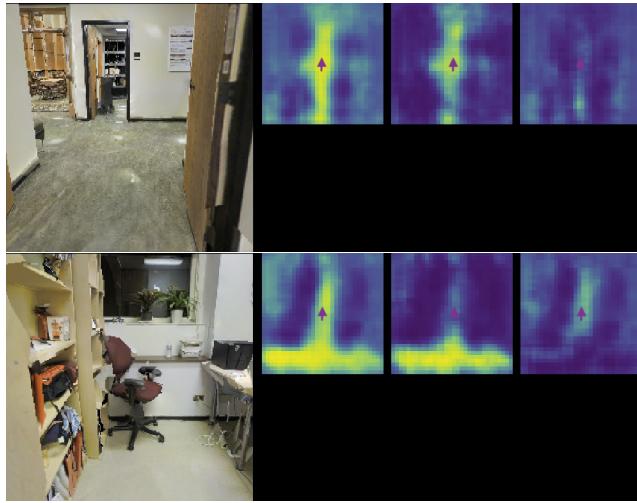


Fig. 9: We visualize the first person image, prediction for all free space, prediction for free space in a hallway, and prediction for free space inside a room (in order). Once again, the predictions are in an egocentric coordinate frame (agent (denoted by the purple arrow) is at the center and looking upwards). The top figure pane shows the case when the agent is actually in a hallway, while the bottom figure pane shows the case when the agent is inside a room.

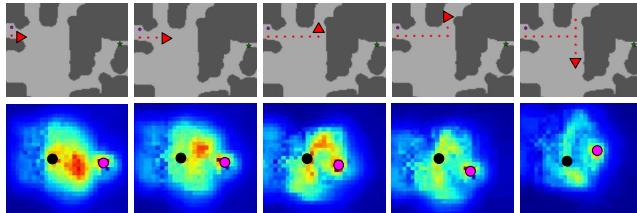


Fig. 10: We visualize the value function for five snapshots for an episode for the single scale version of our model. The top row shows the agent's location and orientation with a red triangle, nodes that the agent has visited with red dots and the goal location with the green star. Bottom row shows a 1 channel projection of the value maps (obtained by taking the channel wise max) and visualizes the agent location by the black dot and the goal location by the pink dot. Initially the agent plans to go straight ahead, as it sees the wall it develops an inclination to turn left. It then turns into the room (center figure), planning to go up and around to the goal but as it turns again it realizes that that path is blocked (center right figure). At this point the value function changes (the connection to the goal through the top room becomes weaker) and the agent approaches the goal via the downward path.

space prediction for two cases where the agent is looking through a doorway. We see that the mapper representation is expressive enough to make reasonable predictions for free space behind the doorway. **This is something that a purely geometric system that only reasons about directly visible parts of the environment is simply incapable of doing.** Finally, we show the output of readout functions that were trained for differentiating between free space in a hallway *vs.* free space in a room. Figure 9 (top) shows the prediction for when the agent is out in the hallway, and Figure 9 (bottom) shows the prediction for when the agent is in a room. We see that the representation produced by the mapper can reasonably

distinguish between free space in a hallway *vs.* free space in a room, even though it was never explicitly trained to do so. Once again, this is something that a purely geometric description of the world will be unable to capture.

Do the value maps obtained from the value iteration module capture the behaviour exhibited by the agent?

Finally, Figure 10 visualizes a one channel projection of the value map for the single scale version of our model at five time steps from an episode. We can see that the value map is indicative of the current actions that the agent takes, and how the value maps change as the agent discovers that the previously hypothesised path was infeasible.

8 Real World Deployment

We have also deployed these learned policies on a real robot. We describe the robot setup, implementation details and our results below.

Robot description. We conducted our experiments on a TurtleBot 2 robot. TurtleBot 2 is a differential drive platform based on the Yujin Kobuki Base. We mounted an Orbbec Astra camera at a height of 80cm, and a GPU-equipped high-end gaming laptop (Gigabyte Aero 15" with an NVIDIA 1060 GPU). The robot is shown in Figure 11 (left). We used ROS to interface with the robot and the camera. We read out images from the camera, and an estimate of the robot's 2D position and orientation obtained from wheel encoders and an onboard inertial measurement unit (IMU). We controlled the robot by specifying desired linear and angular velocities. These desired velocity commands are internally used to determine the voltage that is applied to the two motors through a proportional integral derivative (PID) controller. Note that TurtleBot 2 is a non-holonomic system. It only moves in the direction it is facing, and its dynamics can be approximated as a Dubins Car.

Implementation of macro-actions. Our policies output macro actions (rotate left or right by 90°, move forward 40cm). Unlike past work [12] that uses human operators to implement such macro-actions for such simulation to real transfer, we implement these macro-actions using an iterative linearquadratic regulator (iLQR) controller [41, 55]. iLQR leverages known system dynamics to output a dynamically feasible local reference trajectory (sequence of states and controls) that can convey the system from a specified initial state to a specified final state (in our case, rotation of 90° or forward motion of 40cm). Additionally, iLQR is a state-space feedback controller. It estimates time-varying feedback matrices, that can adjust the reference controls to compensate for deviations from the reference trajectory (due to mis-match in system dynamics or noise in the environment). These adjusted controls are applied to the robot. More details are provided in Section A5.

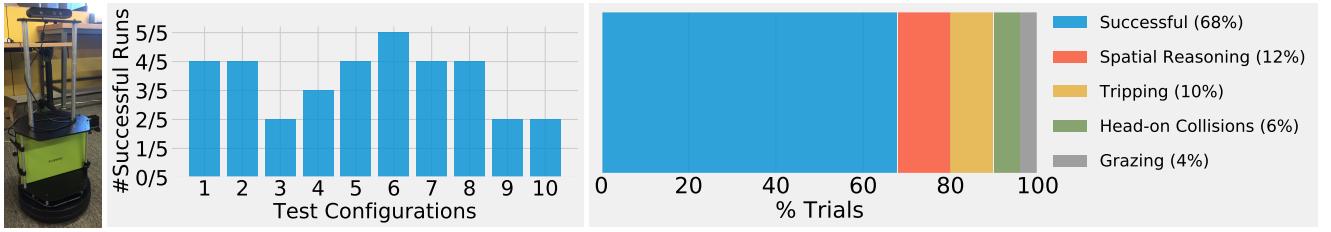


Fig. 11: Real World Deployment: We report success rate on different test cases for real world deployment of our policy on TurtleBot 2. The policy was trained for the geometric task using RGB images in simulation. Right plot shows breakdown of runs. 68% runs succeeded, 20% runs failed due to infractions, and the remaining 12% runs failed as the agent was unable to go around obstacles.

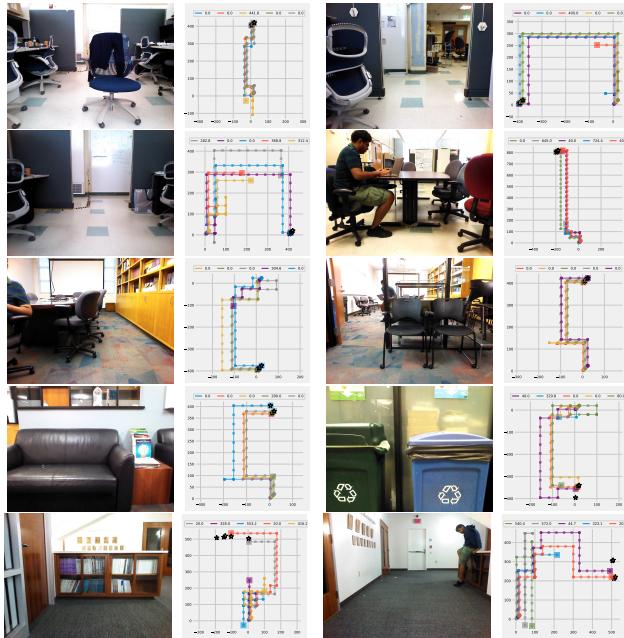


Fig. 12: Real World Experiments: Images and schematic sketch of the executed trajectory for each of the 5 runs for the 10 test cases that were used to test the policy in the real world. Runs are off-set from each other for better visualization. Start location (always (0,0)) is denoted by a solid circle, goal location by a star, and the final location of the agent is denoted by a square. Legend notes the distance of the goal location from the final position. Best seen in color on screen.

Policy. We deployed the policy for the geometric task onto the robot. As all other policies, this policy was trained entirely in simulation. We used the ‘CMP [+6 MP3D Env]’ policy that was trained with the six additional large environments from Matterport3D dataset [14] (on top of the 4 environments from the S3DIS [7] dataset). Apart from improves performance in simulation (SPL from 59.6% to 70.8%), it also exhibited better real world behavior in preliminary runs.

Results. We ran the robot in 10 different test configurations (shown in Figure 12). These tests were picked such that there was no straight path to the goal location, and involved situation like getting out of a room, going from one cubicle to another, and going around tables and kitchen counters. We found the depth as sensed from the Orbbec camera to be very noisy (and different from depth produced in our simulator),

and hence only conducted experiments with RGB images as input. We conducted 5 runs for each of the 10 different test configurations, and report the success rate for the 10 configurations in Figure 11 (middle). A run was considered successful if the robot made it close to the specified target location (within 80cm) without brushing against or colliding with any objects. Sample videos of execution are available on the project website. The policy achieved a success rate of 68%. Executed trajectories are plotted in Figure 12. This is a very encouraging result, given that the policy was trained entirely in simulation on very different buildings, and the lack of any form of domain adaptation. Our robot, that only uses monocular RGB images, successfully avoids running into obstacles and arrives at the goal location for a number of test cases.

Figure 11 (right) presents failure modes of our runs. 10 of the 16 failures are due to infractions (head-on collisions, grazing against objects, and tripping over rods on the floor). These failures can possibly be mitigated by use of a finer action space for more dexterous motion, additional instrumentation such as near range obstacle detection, or coupling with a collision avoidance system. The remaining 6 failures correspond to not going around obstacles, possibly due to inaccurate perception.

9 Discussion

In this paper, we introduced a novel end-to-end neural architecture for navigation in novel environments. Our architecture learns to map from first-person viewpoints and uses a planner with the learned map to plan actions for navigating to different goals in the environment. Our experiments demonstrate that such an approach outperforms other direct methods which do not use explicit mapping and planning modules. While our work represents exciting progress towards problems which have not been looked at from a learning perspective, a lot more needs to be done for solving the problem of goal oriented visual navigation in novel environments.

A central limitation in our work is the assumption of perfect odometry. Robots operating in the real world do not have perfect odometry and a model that factors in uncertainty

in movement is essential before such a model can be deployed in the real world.

A related limitation is that of building and maintaining metric representations of space. This does not scale well for large environments. We overcome this by using a multi-scale representation for space. Though this allows us to study larger environments, in general it makes planning more approximate given lower resolution in the coarser scales which could lead to loss in connectivity information. Investigating representations for spaces which do not suffer from such limitations is important future work.

In this work, we have exclusively used DAGGER for training our agents. Though this resulted in good results, it suffers from the issue that the optimal policy under an expert may be unfeasible under the information that the agent currently has. Incorporating this in learning through guided policy search or reinforcement learning may lead to better performance specially for the semantic task.

Acknowledgments: We thank Shubham Tulsiani, David Fouhey, Somil Bansal and Christian Häne for useful discussion and feedback on the manuscript, as well as Marek Fiser and Sergio Guadarrama for help with the Google infrastructure and Tensorflow.

References

1. Matterport. <https://matterport.com/> ⁴
2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL <http://tensorflow.org/> ⁷ Software available from tensorflow.org
3. Abel, D., Agarwal, A., Diaz, F., Krishnamurthy, A., Schapire, R.E.: Exploratory gradient boosting for reinforcement learning in complex domains. arXiv preprint arXiv:1603.04119 (2016) ³
4. Ammirato, P., Poirson, P., Park, E., Kosecka, J., Berg, A.C.: A dataset for developing and benchmarking active vision. In: ICRA (2017) ⁴
5. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., Zamir, A.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018) ⁷
6. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Hengel, A.v.d.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018) ⁴
7. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: CVPR (2016) ^{4, 7, 8, 14}
8. Aydemir, A., Pronobis, A., Göbelbecker, M., Jensfelt, P.: Active visual object search in unknown environments using uncertain semantics. IEEE Transactions on Robotics (2013) ³
9. Bellman, R.: A markovian decision process. Tech. rep., DTIC Document (1957) ⁵
10. Bhatti, S., Desmaison, A., Miksic, O., Nardelli, N., Siddharth, N., Torr, P.H.: Playing doom with slam-augmented deep reinforcement learning. arXiv preprint arXiv:1612.00380 (2016) ⁴
11. Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J.Z., Rae, J., Wierstra, D., Hassabis, D.: Model-free episodic control. arXiv preprint arXiv:1606.04460 (2016) ³
12. Bruce, J., Sünderhauf, N., Mirowski, P., Hadsell, R., Milford, M.: Learning deployable navigation policies at kilometer scale from a single traversal. arXiv preprint arXiv:1807.05211 (2018) ^{4, 13}
13. Canny, J.: The complexity of robot motion planning. MIT press (1988) ³
14. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017) ^{4, 10, 11, 13}
15. Chaplot, D.S., Parisotto, E., Salakhutdinov, R.: Active neural localization. In: ICLR (2018) ⁴
16. Daftary, S., Bagnell, J.A., Hebert, M.: Learning transferable policies for monocular reactive mav control. In: ISER (2016) ³
17. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: CVPR (2017) ⁴
18. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: CVPR (2018) ⁴
19. Davison, A.J., Murray, D.W.: Mobile robot localisation using active vision. In: ECCV (1998) ²
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009) ⁷
21. Duan, Y., Schulman, J., Chen, X., Bartlett, P.L., Sutskever, I., Abbeel, P.: RL²: Fast reinforcement learning via slow reinforcement learning. arXiv preprint arXiv:1611.02779 (2016) ⁴
22. Elfes, A.: Sonar-based real-world mapping and navigation. RA (1987) ³
23. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. Computer (1989) ³
24. Fraundorfer, F., Heng, L., Honegger, D., Lee, G.H., Meier, L., Tanskanen, P., Pollefeys, M.: Vision-based autonomous mapping and exploration using a quadrotor mav. In: IROS (2012) ³
25. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artificial Intelligence Review (2015) ³
26. Giusti, A., Guzzi, J., Cireşan, D.C., He, F.L., Rodríguez, J.P., Fontana, F., Faessler, M., Forster, C., Schmidhuber, J., Di Caro, G., et al.: A machine learning approach to visual perception of forest trails for mobile robots. RAL (2016) ³
27. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: IQA: Visual question answering in interactive environments. In: CVPR (2018) ⁴
28. Gu, S., Lillicrap, T., Sutskever, I., Levine, S.: Continuous deep q-learning with model-based acceleration. In: ICML (2016) ³
29. Gupta, S., Arbeláez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: CVPR (2013) ³
30. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: CVPR (2016) ^{7, 17}
31. Haarnoja, T., Ajay, A., Levine, S., Abbeel, P.: Backprop KF: Learning discriminative deterministic state estimators. In: NIPS (2016) ⁴
32. Hadsell, R., Sermanet, P., Ben, J., Erkan, A., Scoffier, M., Kavukcuoglu, K., Muller, U., LeCun, Y.: Learning long-range vision for autonomous off-road driving. Journal of Field Robotics (2009) ³
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) ⁵
34. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016) ⁷

35. Heess, N., Hunt, J.J., Lillicrap, T.P., Silver, D.: Memory-based control with recurrent neural networks. arXiv preprint arXiv:1512.04455 (2015) [3](#)
36. Henriques, J.F., Vedaldi, A.: Mapnet: An allocentric spatial memory for mapping environments. In: CVPR (2018) [4](#)
37. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In: ISER (2010) [3](#)
38. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation (1997) [5](#)
39. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015) [7](#)
40. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. UIST (2011) [3](#)
41. Jacobson, D.H., Mayne, D.Q.: Differential dynamic programming (1970) [13](#)
42. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS (2015) [5](#)
43. Kahn, G., Zhang, T., Levine, S., Abbeel, P.: Plato: Policy learning using adaptive trajectory optimization. In: ICRA (2017) [3](#)
44. Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. RA (1996) [3](#)
45. Khan, A., Zhang, C., Atanasov, N., Karydis, K., Kumar, V., Lee, D.D.: Memory augmented control networks. In: ICLR (2018) [4](#)
46. Khatib, O.: Real-time obstacle avoidance for manipulators and mobile robots. IJRR (1986) [3](#)
47. Kim, H., Jordan, M.I., Sastry, S., Ng, A.Y.: Autonomous helicopter flight via reinforcement learning. In: NIPS (2003) [3](#)
48. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [7](#)
49. Kohl, N., Stone, P.: Policy gradient reinforcement learning for fast quadrupedal locomotion. In: ICRA (2004) [3](#)
50. Konolige, K., Bowman, J., Chen, J., Mihelich, P., Calonder, M., Lepetit, V., Fua, P.: View-based maps. IJRR (2010) [3](#)
51. Koppula, H., Anand, A., Joachims, T., Saxena, A.: Semantic labeling of 3d point clouds for indoor scenes. In: NIPS (2011) [3](#)
52. Kuipers, B., Byun, Y.T.: A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. Robotics and autonomous systems (1991) [3](#)
53. Lavalle, S.M., Kuffner Jr, J.J.: Rapidly-exploring random trees: Progress and prospects. In: Algorithmic and Computational Robotics: New Directions (2000) [3](#)
54. Levine, S., Finn, C., Darrell, T., Abbeel, P.: End-to-end training of deep visuomotor policies. JMLR (2016) [2](#)
55. Li, W., Todorov, E.: Iterative linear quadratic regulator design for nonlinear biological movement systems. In: ICINCO (2004) [13](#)
56. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., et al.: Learning to navigate in complex environments. In: ICLR (2017) [4](#)
57. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T.P., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: ICML (2016) [3](#)
58. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature (2015) [2, 3](#)
59. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: CVPR (2004) [4](#)
60. Oh, J., Chockalingam, V., Singh, S., Lee, H.: Control of memory, active perception, and action in minecraft. In: ICML (2016) [3](#)
61. Parisotto, E., Salakhutdinov, R.: Neural map: Structured memory for deep reinforcement learning. In: ICLR (2018) [4](#)
62. Peters, J., Schaal, S.: Reinforcement learning of motor skills with policy gradients. Neural networks (2008) [3](#)
63. Ross, S., Gordon, G.J., Bagnell, D.: A reduction of imitation learning and structured prediction to no-regret online learning. In: AISTATS (2011) [6, 7](#)
64. Sadeghi, F., Levine, S.: (CAD)²RL: Real singel-image flight without a singel real image. In: RSS (2017) [4](#)
65. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. In: ICLR (2018) [4](#)
66. Savva, M., Chang, A.X., Dosovitskiy, A., Funkhouser, T., Koltun, V.: MINOS: Multimodal indoor simulator for navigation in complex environments. arXiv:1712.03931 (2017) [4](#)
67. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) [8](#)
68. Schulman, J., Levine, S., Moritz, P., Jordan, M.I., Abbeel, P.: Trust region policy optimization. In: ICML (2015) [3](#)
69. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. IJCV (2008) [3](#)
70. Tamar, A., Levine, S., Abbeel, P.: Value iteration networks. In: NIPS (2016) [3, 5, 6](#)
71. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics. MIT press (2005) [2, 3](#)
72. Tolman, E.C.: Cognitive maps in rats and men. Psychological review **55**(4), 189 (1948) [2](#)
73. Toussaint, M.: Learning a world model and planning with a self-organizing, dynamic neural system. In: NIPS (2003) [3](#)
74. Wierstra, D., Förster, A., Peters, J., Schmidhuber, J.: Recurrent policy gradients. Logic Journal of IGPL (2010) [3](#)
75. Wu, Y., Wu, Y., Gkioxari, G., Tian, Y.: Building generalizable agents with a realistic and rich 3d environment. arXiv preprint arXiv:1801.02209 (2018) [4](#)
76. Xia, F., R. Zamir, A., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: CVPR (2018) [4](#)
77. Yamauchi, B.: A frontier-based approach for autonomous exploration. In: Computational Intelligence in Robotics and Automation, 1997. CIRA'97., Proceedings., 1997 IEEE International Symposium on (1997) [8](#)
78. Zamir, A.R., Wekel, T., Agrawal, P., Wei, C., Malik, J., Savarese, S.: Generic 3d representation via pose estimation and matching. In: ECCV (2016) [3](#)
79. Zhang, M., McCarthy, Z., Finn, C., Levine, S., Abbeel, P.: Learning deep neural network policies with continuous memory states. In: ICRA (2016) [3](#)
80. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: ICRA (2017) [2, 3, 8](#)

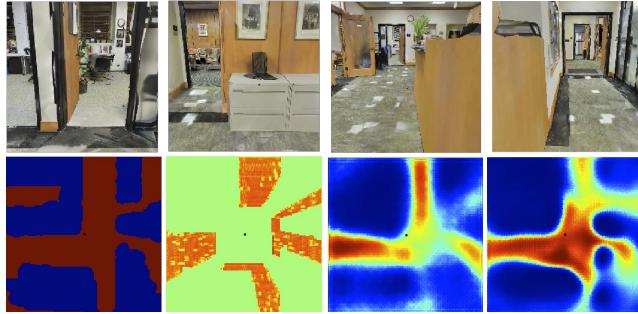


Fig. A1: Output Visualization for Mapper trained for Free Space Prediction: We visualize the output of the mapper when directly trained for task of predicting free space. We consider the scenario of an agent rotating about its current position, the task is to predict free space in a 3.20 meter neighborhood of the agent, supervision for this experiment at end of the agents rotation. The top row shows the 4 input views. The bottom row shows the ground truth free space, predicted free space by analytically projecting the depth images, learned predictor using RGB images and learned predictor using depth images. Note that the learned approaches produce more complete output and are able to make predictions where no observations were made.

A1 Backward Flow Field ρ from Egomotion

Consider a robot that rotates about its position by an angle θ and then moves t units forward. Corresponding points p in the original top-view and p' in the new top-view are related to each other as follows (R_θ is a rotation matrix that rotates a point by an angle θ):

$$p' = R_\theta p - t \text{ or } p = R_\theta(p' + t) \quad (3)$$

Thus given the egomotion θ and t , for each point in the new top-view we can compute the location in the original top-view from which it came from.

A2 Mapper Performance in Isolation

To demonstrate that our proposed mapper architecture works we test it in isolation on the task of free space prediction. We consider the scenario of an agent rotating about its current position, and the task is to predict free space in a 3.20 meter neighborhood of the agent. We only provide supervision for this experiment at end of the agents rotation. Figure A1 illustrates what the mapper learns. Observe that our mapper is able to make predictions where no observations are made. We also report the mean average precision for various versions of the mapper Table A1 on the test set (consisting of 2000 locations from the testing environment). We compare against an analytic mapping baseline which projects points observed in the depth image into the top view (by back projecting them into space and rotating them into the top-down view).

A3 Additional Experiments

Additional experiment on an internal Matterport dataset. We also conduct experiments on an internal Matterport dataset consisting of 41 scanned environments. We train on 27 of these environments, use 4 for validation and test on the remaining 10. We show results for the 10 test environments in Figure A2. We again observe that CMP consistently outperforms the 4 frame reactive baseline and LSTM.

Ablations. We also present performance of ablated versions of our proposed method in Table A2.

Method	Modality	CNN Architecture	Free Space Prediction AP
Analytic Projection	depth	-	56.1
Learned Mapper	RGB	ResNet-50	74.9
Learned Mapper	depth	ResNet-50 Random Init.	63.4
Learned Mapper	depth	ResNet-50 Init. using [30]	78.4

Table A1: Mapper Unit Test: We report average precision for free space prediction when our proposed mapper architecture is trained directly for the task of free space prediction on a test set (consisting of 2000 locations from the testing environment). We compare against an analytic mapping baseline which projects points observed in the depth image into the top view (by back projecting them into space and rotating them into the top-down view).

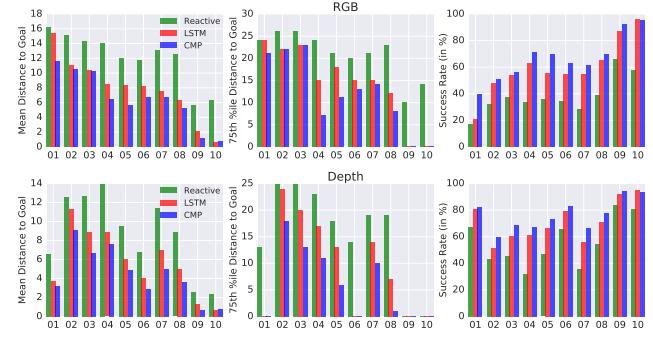


Fig. A2: We report the mean distance to goal, 75th percentile distance to goal (lower is better) and success rate (higher is better) for Reactive, LSTM and CMP based agents on different test environments from an internal dataset of Matterport scans. We show performance when using RGB images (top row) and depth images (bottom row) as input. We note that CMP consistently outperforms Reactive and LSTM based agents.

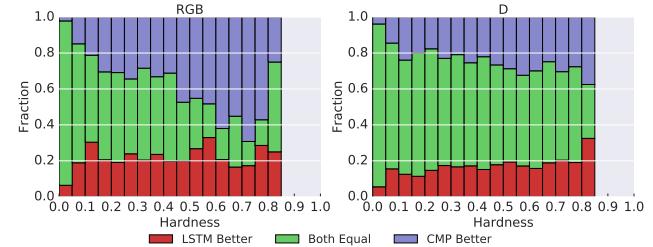


Fig. A3: We show how performance of LSTM and CMP compare across geometric navigation tasks of different hardness. We define hardness as the gap between the ground truth and heuristic (Manhattan) distance between the start and goal, normalized by the ground truth distance. For each range of hardness we show the fraction of cases where LSTM gets closer to the goal (LSTM Better), both LSTM and CMP are equally far from the goal (Both Equal) and CMP gets closer to goal than LSTM (CMP Better). We show results when using RGB images as input (left plot) and when using Depth images as input (right plot). We observe that CMP is generally better across all values of hardness, but for RGB images it is particularly better for cases with high hardness.

Method	Mean		75 th %ile		Success %age	
	RGB	Depth	RGB	Depth	RGB	Depth
Geometric Task						
Initial	25.3	25.3	30	30	0.7	0.7
No Image LSTM	20.8	20.8	28	28	6.2	6.2
CMP						
Full model	7.7	4.8	14	1	62.5	78.3
Single-scale planner	7.9	4.9	12	1	63.0	79.5
Shallow planner	8.5	4.8	16	1	58.6	79.0
Analytic map	-	8.0	-	14	-	62.9

Table A2: Ablative Analysis for CMP: We follow the same experimental setup as used for table in the main text. See text for details.

	Mean			75 th %ile			Success Rate (in %)		
	Init.	LSTM	CMP	Init.	LSTM	CMP	Init.	LSTM	CMP
Far away goal (maximum 64 steps away)									
Run for 79 steps	47.2	15.2	11.9	58	29	19.2	0.0	58.4	66.3
Run for 159 steps	47.2	12.5	9.3	58	19	0	0.0	69.0	78.5
Generalization									
Train on 1 floor	25.3	8.9	7.0	30	18	10	0.7	58.9	67.9
Transfer from IMD	25.3	11.0	8.5	30	21	15	0.7	48.6	61.1

Table A3: We report additional comparison between best performing models. See text for details.

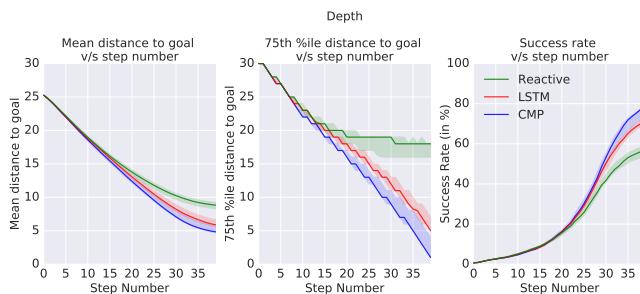


Fig. A4: We show the variance in performance over five re-trainings from different random initializations of the agents when using depth images as input (the solid line plots the median performance and the surrounding shaded region represents the minimum and maximum value across five different runs). We note that the variation in performance is reasonably small for all models and CMP consistently outperforms the two baseline.

Single Scale Planning. We replace the multi-scale planner with a single-scale planner. This results in slightly better performance but comes at the cost of increased planning cost.

No Planning. We swap out the planner CNN with a shallower CNN. This also results in drop in performance specially for the RGB case as compared to the full system which uses the full planner.

Analytic Mapper. We also train a model where we replace our learned mapper for an analytic mapper that projects points from the depth image into the overhead view and use it with a single scale version of the planner. We observe that this analytic mapper actually works worse than the learned one thereby validating our architectural choice of learning to map.

Additional comparisons between LSTM and CMP. We also report additional experiments on the Stanford S3DIS dataset to further compare the performance of the LSTM baseline with our model in the most competitive scenario where both methods use depth images. These are reported in Table A3. We first evaluate how well do these models

perform in the setting when the target is much further away (instead of sampling problems where the goal is within 32 time steps we sample problems where the goal is 64 times steps away). We present evaluations for two cases, when this agent is run for 79 steps or 159 steps (see ‘Far away goal’ rows in Table A3). We find that both methods suffer when running for 79 steps only, because of limited time available for back-tracking, and performance improves when running these agents for longer. We also see a larger gap in performance between LSTM and CMP for both these test scenarios, thereby highlighting the benefit of our mapping and planning architecture.

We also evaluate how well these models generalize when trained on a single scene (‘Train on 1 scene’). We find that there is a smaller drop in performance for CMP as compared to LSTM. We also found CMP to transfer from internal Matterport dataset to the Stanford S3DIS Dataset slightly better (‘Transfer from internal dataset’).

We also study how performance of LSTM and CMP compares across geometric navigation tasks of different hardness in Figure A3. We define hardness as the gap between the ground truth and heuristic (Manhattan) distance between the start and goal, normalized by the ground truth distance. For each range of hardness we show the fraction of cases where LSTM gets closer to the goal (LSTM Better), both LSTM and CMP are equally far from the goal (Both Equal) and CMP gets closer to goal than LSTM (CMP Better). We observe that CMP is generally better across all values of hardness, but for RGB images it is particularly better for cases with high hardness.

A4 Simulation Testbed Details

We pre-processed the meshes to compute space traversable by the robot. Top views of the obtained traversable space are shown in Figures A5 and A6 (training and validation) and Figure A7 (testing) and indicate the complexity of the environments we are working with and the differences in layouts between the training and testing environments. Recall that robot’s action space $\mathcal{A}_{x,\theta}$ consists of macro-actions. We pick θ to be $\pi/2$ which allows us to pre-compute the set of locations (spatial location and orientation) that the robot can visit in this traversable space. We also precompute a directed graph $\mathcal{G}_{x,\theta}$ consisting of this set of locations as nodes and a connectivity structure based on the actions available to the robot.

Our setup allows us to study navigation but also enables us to independently develop and design our mapper and planner architectures. We developed our mapper by studying the problem of free space prediction from sequence of first person view as available while walking through these environments. We developed our planner by using the ground truth top view free space as 2D mazes to plan paths through. Note that this division was merely done to better understand each component, the final mapper and planner are trained jointly and there is no restriction on what information gets passed between the mapper and the planner.

A5 Macro-action Implementation using ILQR

We use the robot 2D location and orientation as the state s , the linear and angular velocity as the control inputs u to the system, and function f to model the dynamics of the system as follows:

$$\mathbf{s}_t = \begin{bmatrix} x_t \\ y_t \\ \omega_t \end{bmatrix} \quad \mathbf{u}_t = \begin{bmatrix} v_t \\ \omega_t \end{bmatrix} \quad f(\mathbf{s}_t, \mathbf{u}_t) = \begin{bmatrix} x_t + v_t \Delta t \cos(\theta_t) \\ y_t + v_t \Delta t \sin(\theta_t) \\ \theta_t + \omega_t \Delta t \end{bmatrix} \quad (4)$$

Given an initial state \mathbf{s}_0 , and a desired final state \mathbf{s}_T ($= \mathbf{0}$ without loss of generality), iLQR solves the following optimization problem:

$$\arg \min_{\mathbf{u}_t} \quad \sum_t \mathbf{s}_t^T Q \mathbf{s}_t + \mathbf{u}_t^T R \mathbf{u}_t \quad (5)$$

$$\text{subject to } \mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{u}_t) \text{ for } t \in [1, \dots, T] \quad (6)$$

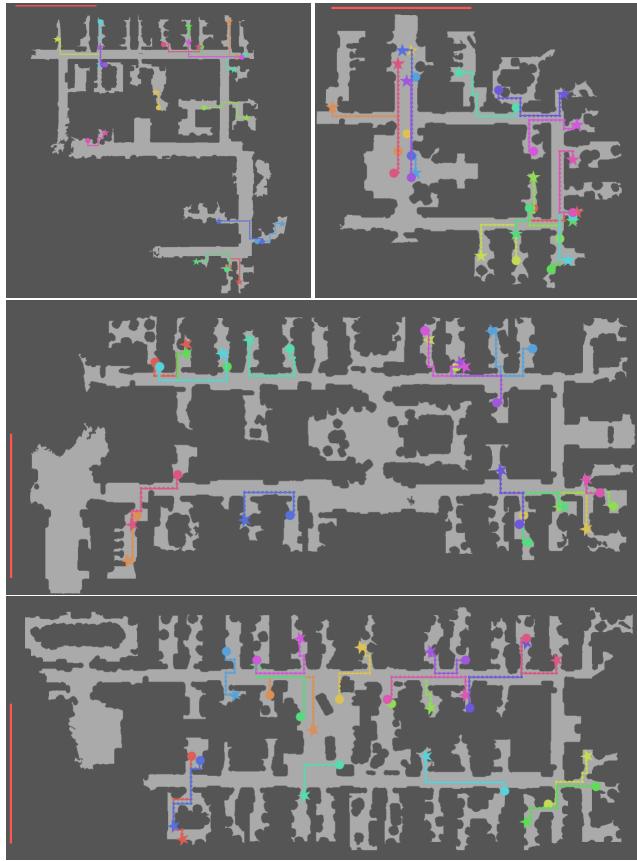


Fig. A5: Maps for *area52*, *area3* and *area6*. Light area shows traversable space. Red bar in the corner denotes a length of 32 units (12.80 metres). We also show some example geometric navigation problems in these environments, the task is to go from the circle node to the star node.

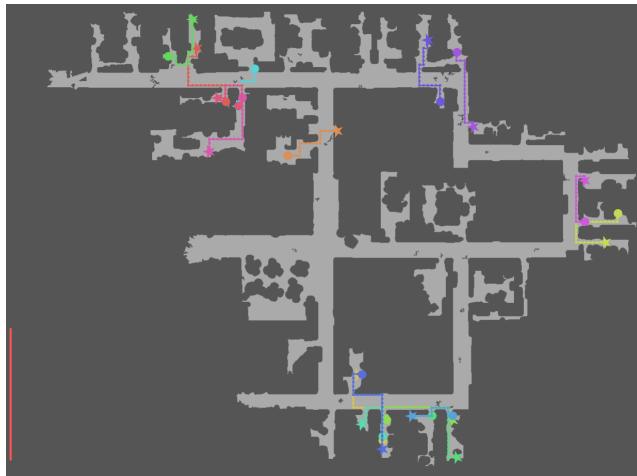


Fig. A6: Map for *area51*. Light area shows traversable space. Red bar in the corner denotes a length of 32 units (12.80 metres). We also show some example geometric navigation problems in these environments, the task is to go from the circle node to the star node.

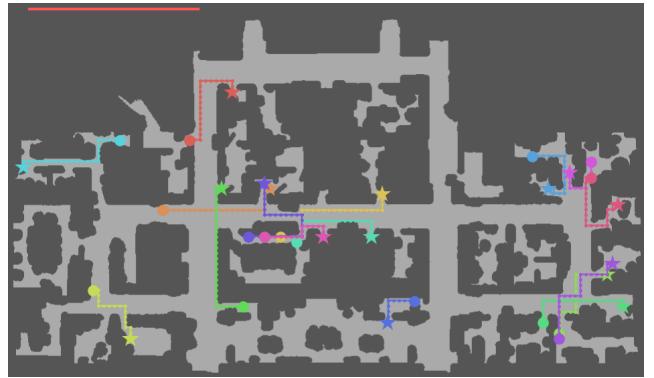


Fig. A7: Map for *area4*. This floor was used for testing all the models. Light area shows traversable space. Red bar in the corner denotes a length of 32 units (12.80 metres). We also show some example geometric navigation problems in these environments, the task is to go from the circle node to the star node.

where, matrices Q and R are specified to be appropriately scaled identity matrices, Δt controls the frequency with which we apply the control input, and T determines the total time duration we have to finish executing the macro-action. Matrix Q incentivizes the system to reach the target state quickly, and matrix R incentivizes applying small velocities. The exact scaling of matrices Q and R , Δt and T are set experimentally by running the robot on a variety of start and goal state pairs.

Given Dubins Car dynamics are non-linear, iLQR optimizes the cost function by iteratively linearizing the system around the current solution. As mentioned, iLQR outputs $\mathbf{x}_t^{\text{ref}}$, $\mathbf{u}_t^{\text{ref}}$, and a set of feedback matrices K_t . The control to be applied to the system at time step t is obtained as $\mathbf{u}_t^{\text{ref}} + K_t (\tilde{\mathbf{s}}_t - \mathbf{s}_t^{\text{ref}})$, where $\tilde{\mathbf{s}}_t$ is the estimated state of the system as measured from the robots wheel encoders and IMU (after appropriate coordinate transforms).

A6 Change Log

- v1.** First Version. **v2.** CVPR 2017 Camera Ready Version. Added more experiments for semantic task. Code made available on project website.
- v3.** IJCV Version. Putting work in context of more recent work in the area, added SPL metric, added comparisons to classical methods, added details about real world deployment of the learned policies, added more visualizations of what is being learned inside the network.