Learning Your Way Without Map or Compass: Panoramic Target Driven Visual Navigation

David Watkins-Valls∗, Jingxi Xu∗, Nicholas Waytowich2 and Peter Allen1

Abstract—We present a robot navigation system that uses an imitation learning framework to successfully navigate in complex environments. Our framework takes a pre-built 3D scan of a real environment and trains an agent from pre-generated expert trajectories to navigate to any position given a panoramic view of the goal and the current visual input without relying on map, compass, odometry, GPS or relative position of the target at runtime. Our end-to-end trained agent uses RGB and depth (RGBD) information and can handle large environments (up to $10^3 m^2$) across multiple rooms (up to 40) and generalizes to unseen targets. We show that when compared to several baselines using deep reinforcement learning and RGBD SLAM, our method (1) requires fewer training examples and less training time, (2) reaches the goal location with higher accuracy, (3) produces better solutions with shorter paths for long-range navigation tasks, and (4) generalizes to unseen environments given an RGBD map of the environment.

I. INTRODUCTION

The ability to navigate efficiently and accurately within an environment is fundamental to intelligent behavior and has been a focus of research in robotics for many years. Traditionally, robotic navigation is solved using model-based methods with an explicit focus on position inference and mapping, such as Simultaneous Localization and Mapping (SLAM) [1]. These models use path planning algorithms, such as Probabilistic Roadmaps (PRM) [2] and Rapidly Exploring Random Trees (RRT) [3], [4] to plan a collision-free path. These methods ignore the rich information from visual input and are highly sensitive to robot odometry and noise in sensor data. For example, a robot navigating through a room may lose track of its position due to the navigation software not properly modeling friction.

Model-free reinforcement learning (RL) agents have performed well on many robotic tasks [5], [6], [7], [8], leading researchers to rely on RL for robotic navigation tasks [9], [10], [11], [12]. Recent work in robotic visual navigation uses reinforcement learning which trains an agent to navigate to a goal using only the current and goal RGB images [9]. While reinforcement learning has the convenience of being weakly supervised, it generally suffers from sparse rewards in navigation, requires a huge number of training episodes to converge, and struggles to generalize to unseen targets. The problem is further exacerbated when the navigation environment becomes large and complex (across multiple rooms and scenes with various obstacles), leading to difficult long-range path solutions.

New advancements in annotated 3D maps of real world data, such as Stanford2D3DS [13] and Matterport3D [14], enable the collection of large amounts of trajectory data. Simulators capable of collecting this data have arisen in the past few years in the form of MINOS [15], Gibson [16], Habitat [17], and THOR [9]. These systems enable simulta-

*Authors have contributed equally and names are in alphabetical order.
1Department of Computer Science, Columbia University, New York, NY, USA. {davidwatkins, allen}@cs.columbia.edu, jingxi.xu@columbia.edu
2U.S. Army Research Laboratory, Baltimore, MD, USA. nicholas.r.waytowich.civ@mail.mil

This work is supported by NSF Grant CMMI 1734557. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-18-2-0244. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The title is an homage to Harold Gatty’s book Finding Your Way Without Map or Compass.
neous use of real and simulated environments for training, without the need for visual domain adaptation. Gibson [16] features real-world and photo-realistic data generated from fully scanned 3D homes and buildings that allow for easier collection of demonstration data for supervised learning.

Our work focuses on exploring supervised methods (in particular, imitation learning) to bring better performance to robotic visual navigation, while taking advantage of the current progress in robotic simulators and datasets to efficiently collect training data and handle the sim-to-real gap and domain adaptation. Our contributions are as follows: (1) we provide a navigation pipeline where the agent learns to navigate to unseen targets using the current RGBD view and a novel 8-image panoramic goal without using GPS, compass, map, or relative position of goals at runtime, (2) we provide a framework to efficiently generate optimal expert trajectories in the Gibson simulator using a 3D scan of the environment of interest, and (3) we provide a methodology for discretizing a continuous trajectory into a series of \{forward, right, left\} commands. Our method outperforms alternative methods in both the quality of the solution paths and the success rate, with significantly fewer training examples and less training time. Videos, dataset and source code can be found at \url{http://crlab.cs.columbia.edu/learning_your_way/}

II. RELATED WORK

a) Reinforcement learning methods for navigation: Previous work in visual navigation [9] provides a target-driven reinforcement learning framework for robotic visual navigation. Our method shares the same objective of navigating to the goal position using the goal image, the current image and a sequence of history images. [18], [17] trains an RL agent to navigate in realistic cluttered environments using a PointGoal (a specific location for the goal target). They assumed an idealized GPS which provides the relative goal position and use this information to train their agents. Both [17] and [9] claim that their learned policy generalizes across targets and environments. [9] only evaluates their method on new targets that are several steps away from the targets that the agent is trained on and the scene-specific layer has to be retrained for the policy to work in a new environment. [17] relies on an idealized GPS and the specific location of the goal and it generalizes to new environments by learning a bug algorithm like behavior to follow the boundaries. Imagine a scenario where a person is placed into a building they have not seen before with nothing but an image of the place they need to get to. It would be unfair for us to expect this person to navigate to the target location in any efficient manner. Therefore, a robotic agent would be unable to handle generalizing to new untrained environments using vision alone and we focus on the ability to generalize to any unseen target in the environment that the robotic agent has seen before.

Previous work [11] uses a synthetic 3D maze environment and the agent is trained on a single goal. Another work [19] trains an agent to navigate using real world Google Map views with the goal coordinate provided. [12], [10], [20] present hierarchical robot navigation methods using reinforcement learning to learn local and short-range obstacle avoidance tasks and using sampling-based path planning algorithms as global planners. These methods use 1D lidar sensor data and a dynamic goal position as input. [21], [22] use value iteration networks [23] to learn navigation strategies in simplistic synthetic simulated environments. [24] evaluates different representations for target-driven visual navigation using a semantic target and an off-the-shelf segmenter. [25] presents a method with an interactive world model to navigate to a fixed goal in a known environment. None of these works solve the problem of indoor visual navigation because they either make compromises in sensory input (1D lidar, goal position) or have a different environmental design (synthetic maps, outdoor data).

b) Supervised learning methods for navigation: Because most work using learning methods for robotic navigation relies on deep RL, supervised methods are less explored. [26], [27] uses Convolutional Neural Networks (CNNs) for robotic navigation. However, their methods are different from ours in that they use odometry and a goal location as input into their networks and their models are trained only for simple tasks such as collision detection and localization.

c) Datasets and simulators for navigation environments: As the broader area of active and embodied perception has received increased interest, new datasets (Stanford2D3DS [13], Matterport3D [14], SUNCG [28], Gibson [16]) and simulators (MINOS [15], Gibson [16], Habitat [17], AI2THOR [29]) have been created for robotic navigation. These new environments enable researchers to train an agent in simulation using real world data and obtain training data much faster than would be possible in the real world alone. They allow agents to be trained at scale using ground truth positioning and fast rendering. The Gibson simulator uses PyBullet [30] to simulate collisions with the environment as well as dynamic environment tasks.

III. PROBLEM SETUP

The goal of this work is to enable the robot to autonomously navigate to a target position, described by a set of panoramic images taken at the goal, without providing any odometry, GPS or relative location of the target but only RGBD input from the robot’s point of view. We do not require a particular orientation at the target location. The problem is referred to as \textit{target-driven visual navigation} in the literature [9], where the task objective (i.e., navigation destination) is specified as inputs to our model. Traditional learning-based visual navigation methods have largely focused on learning goal-specific models that tackle individual tasks in isolation, where the goal information is hard-coded in the neural network representations, leading to poor generalization to unseen/unexplored targets. Target-driven approaches learn to navigate to new targets without re-training, using a single navigation pipeline.

Our navigation pipeline, denoted as $\Pi$, takes as input the observation of the current state $s_t$ at time step...
require human intervention through the whole process. It learns to stop at the goal and does not need to explicitly check if the goal has been achieved. Our navigation pipeline consists of three separately trained models using neural networks, the autoencoder model $A$, the policy model $E$ and the goal checking model $G$.

The autoencoder model generates latent representations (i.e., embeddings) for both the current state $s_t$ and the goal $g$, denoted $A(s_t)$ and $A(g)$. The policy model takes as input the embeddings of the current state and the panoramic goal and produces a probability distribution over three actions, $a_t \in \{\text{forward, right, left}\} \sim E(A(s_t), A(g))$. It then picks the action with highest probability from this distribution. The goal checking model is responsible for leading the agent towards the goal with as little exploration as possible. The goal checking model is a binary function which takes the same input as the policy model, and decides if the agent has reached the target or not, denoted by $G(A(s_t), A(g)) \in \{0, 1\}$, where 1 corresponds to done and 0 corresponds to not done. An overview of the navigation pipeline is shown in Figure 2.

### B. Autoencoder Model

The autoencoder is able to compress a $256 \times 256 \times 4$ RGBD image into the 4096-d latent space $(\times 64$ space savings). It is then used to encode each image of the state and each image of the panoramic goal. A detailed topology of the encoder section is pictured in Figure 4a. An example of the autoencoder performance is shown in Figure 3.

### C. Policy Model

The policy model takes as input the embeddings of stacked observations and the panoramic goal images to generate the next action $a_t \in \{\text{forward, right, left}\}$. We also evaluated the performance of a variety of other deep learning architectures including convolution along the temporal dimension and long short-term memory (LSTM) [34], with different numbers of past images in the state and different number of panoramic goal images, but the MLP architecture with 4 past images and 1 current image outperforms the others. Its larger number of parameters increases its ability to model complex functions.

The embeddings of the state and the panoramic goal are first concatenated to form a $13 \times 4096$ matrix and then progress through 3 fully-connected layers followed by batch normalization and ReLU activation after each layer, to generate a 16-d vector. The 16-d vector passes through the last fully-connected layer to generate 3 logits. Softmax activation then outputs a distribution over 3 actions $\{\text{forward, right, left}\}$. We use Adam optimizer on the cross-entropy loss for back propagation. At testing, we pick the action with highest probability deterministically.

### IV. NAVIGATION PIPELINE

#### A. Overview

Our navigation pipeline $\Pi$ consists of three separately trained models using neural networks, the autoencoder model $A$, the policy model $E$ and the goal checking model $G$.

The autoencoder model generates latent representations (i.e., embeddings) for both the current state $s_t$ and the goal $g$, denoted $A(s_t)$ and $A(g)$. The policy model takes two inputs, the embeddings of the current state and the embeddings of the target, and produces a probability distribution over three actions, $a_t \in \{\text{forward, right, left}\} \sim E(A(s_t), A(g))$. It then picks the action with highest probability from this distribution. The policy model is responsible for leading the agent towards the goal with as little exploration as possible. The goal checking model is a binary function which takes the same input as the policy model, and decides if the agent has reached the target or not, denoted by $G(A(s_t), A(g)) \in \{0, 1\}$, where 1 corresponds to done and 0 corresponds to not done. An overview of the navigation pipeline is shown in Figure 2.

#### B. Autoencoder Model

Because the input into our neural network models is RGBD images, the training is more efficient if we use embeddings of the raw input. Instead of extracting features from an intermediate layer of a pre-trained classifier such as ResNet-50 [9], [31], we train an autoencoder from images captured from the same environment. Similar to RedNet [32], our autoencoder network is based on a 6-layer CNN with batch normalization on every layer. The reconstruction half of the network is made up of an additional 6 transposed convolutional layers with batch normalization applied before each transposed convolution. We use rectified linear unit (ReLU) as the activation function and Adam optimizer [33] to minimize the mean squared error between the reconstructed and the original images. The autoencoder performance is shown in Figure 3.
runtime, in order to reduce noise, the agent does a 360-degree cross-entropy loss. While using the learned goal checker at runtime, the agent outputs done only if the average probability is over 0.9.

V. Experiments

We evaluate our navigation pipeline in 2 environments selected from the Stanford2D3DS dataset and 3 environments selected from the Matterport3D dataset. Metadata (including number of rooms and area) are shown in Table I. We use the Gibson simulator with a Fetch [35] robot and focus on how the navigation pipeline generalizes to unseen targets under the same trained environment. All experiments are conducted using an NVIDIA 1080Ti GPU. Examples of planned paths, recovery behavior, and experimental environments can be found in the attached video.

A. Network Training Setup

For each submodule (autoencoder, policy, and goal checker) in our navigation pipeline, we need to generate the corresponding dataset for training and testing using the colored meshes loaded in the Gibson simulator. We pick the model with least loss (autoencoder) or highest accuracy (policy, goal checker) on the test set. Unless otherwise specified, we use a learning rate of 0.001.

1) 2D Environment Map: We create a high resolution 2D map of the environment from its colored mesh, which enables trajectory planning using Dijkstra’s algorithm. Each of the maps uses a resolution of 1cm x 1cm per pixel which allows for diverse images from many candidate locations. We check collision for each candidate location using a bounding box with dimensions matching the Fetch robot.

2) Autoencoder: For each environments’ autoencoder dataset, we randomly sample 120K collision-free locations from the generated map and capture RGBD images at those locations in the simulator. We use a 0.9/0.1 train/test split and the autoencoder model is trained for 200 epochs which on average takes 12 GPU hours.

3) Policy: Using Dijkstra’s algorithm with costmap optimization, we are able to generate collision-free expert trajectories. In order to map the trajectory into a sequence of \{forward, right, left\} commands, we step through each waypoint position and use a simple discretization strategy. We look ahead 25 steps to determine whether to turn the robot left or right. The robot turns left or right whether the look-ahead position is turned at an offset greater than 20° and outputs turns in increments of 10°. Using only the next position would result in the robot constantly recalculating its direction and the robot would turn at every step to face a new direction. Forward commands are given if the robot is farther in the next position than 0.1m, and if so moves the robot forward in increments of 0.1m until under the threshold.

Once these commands are generated, we execute the trajectory in the Gibson simulator capturing the RGBD view along every step. This results in a large supervised learning dataset of varying trajectory lengths that we could use for training the agent to learn the policy. We use the trained autoencoder to generate embeddings of RGBD images at

D. Goal Checking Model

The goal checking model takes in the embeddings of the current observation concatenated with the panoramic goal images and predicts whether the agent is at the target position, as shown in Figure (a).

This model is created in response to an optimization on our original architecture which had the policy model output a done action when the robotic agent arrives at a goal position. We find that the training data is too sparse for the agent to effectively learn identifying a goal location because we only have one positive example of done at the end of each trajectory. All the other steps are negative examples for not done. There is a significant imbalance in the number of positive and negative examples. In addition, at runtime our robot is likely to arrive at the target position from a different viewpoint than those in the panoramic goal images, but during training the policy model receives a view that is one of the panoramic goal images. Thus, we implement an additional binary classifier to identify whether the agent has arrived at the goal location. When this model predicts a done action the navigation pipeline terminates.

The goal checking model is a dual-branch network with 1D convolution over the panoramic goal branch. The 1024-d vector from the goal-branch is then concatenated with the current embedding branch to form a 5120-d fused vector, which then passes through an MLP with a hidden layer of 512 units to output the probability of the goal being reached. Weights are then updated using an Adam optimizer on the cross-entropy loss. While using the learned goal checker at runtime, in order to reduce noise, the agent does a 360° rotation when its belief of reaching the goal is over 0.99. It calls the learned goal checker after each 10° turn. The agent outputs done only if the average probability is over 0.9.
each step in the expert trajectories. The average number of steps per trajectory varies from 41 (house2) to 332 (areal). Each training/testing example is constructed by taking the embeddings from past 4 steps concatenated with the embedding at the current step. Because larger environments tend to have longer trajectories, we generate 3000 trajectories for areal and area2, 5000 trajectories for house17, and 7000 trajectories for house1 and house2 to keep the total number of training examples roughly the same. We use 80% of the trajectories for training and the rest for testing. The policy network is trained for 200 epochs which takes around 90 GPU hours. The average accuracy for the policy model is 0.91.

4) Goal Checker: For each environment, we collect positive training examples by randomly sampling 150K positions in the environment for the panoramic goal images and then sample another position within a 0.1m radius for the current image. We collect negative examples by randomly sampling 150K positions for the panoramic goal images and then sample another position at least 1m away from the current image. We use a 0.9/0.1 train/test split and train the network for 300 epochs. It takes 36 GPU hours and the average accuracy is 0.95 across all environments.

B. Comparison Methods

We compare our navigation pipeline with an RGBD SLAM [36] approach and the target-driven deep RL method from [9]. The methods we examine are described below.

a) SLAM is the Real-Time Appearance-Based Mapping (RTABMap) library provided by [36], which is an RGBD Graph-Based SLAM approach based on an incremental appearance-based loop closure detector. The robot is not given the map beforehand in our implementation and builds it up as it moves along.

b) Siamese Actor-Critic (SAC) is the method proposed by [9]. We only keep one scene specific layer and train the whole network for each environment as we focus on the generalization to different targets under the same environment. We provide a goal-reaching reward of 10.0 upon task completion and a small penalty of $-0.01$ at each time step. We train it on 100 targets with a maximum step size of 10000 for each episode. We give each environment a budget of 20M frames (steps).

c) Ours (GPS) is a variant of our proposed pipeline. Instead of using the learned goal checking model, it uses the GPS information and the provided goal coordinate to check whether the agent is at the goal.

d) Ours (no GPS) is our proposed pipeline with the learned goal checker, without GPS, odometry or goal coordinate provided.

a), b) and c) assume the agent has an idealized GPS and is provided with the static goal coordinate as in [17]. As a result, the agent is able to compute the relative position of the target at each time step and can use this information to check if the goal has been reached.

C. Evaluation Criteria

We evaluate the performance of our model using 400 randomly sampled start-goal pairs for which we make sure that a valid path exists. The start and goal locations have never appeared in the training examples. We start the agent at the starting position and provide it with 8 panoramic goal images taken at the goal location. The objective is to navigate to the goal position (no requirements on the robot’s final orientation) autonomously with the shortest path possible using only visual input. Our success tolerance is a 0.5m radius within the target position. Unlike previous works which do not penalize collision through training and allow
collision at runtime [9], [17], we simulate real physics and consider the trial a failure when collision occurs.

Similar to many previous works on navigation benchmarks [9], [17], [18], [37], we focus on three evaluation metrics:

1) **Success Rate** is the number of successful trials over the total number of trials.

2) **Success Weighted by Path Length (SPL)** [18] metric is shown in Formula (2) where \( l_i \) is the length of the shortest path between start and goal position, \( p_i \) is the length of the observed path taken by an agent, and \( S_i \) is a binary indicator of success in trial \( i \). This metric weights each success by the quality of path and thus is always \( \leq \) **Success Rate**.

\[
SPL = \frac{1}{N} \sum_{i=1}^{N} \frac{S_i}{\max(p_i, l_i)}
\] (2)

3) **Observed over Optimal Ratio (OOR)** measures the average ratio of observed path length over optimal path length for successful trials.

### D. Navigation Results

Our proposed navigation pipeline significantly outperforms RGBD SLAM and the state-of-the-art deep RL method in terms of path quality and success rate, as shown in Table I. See Figure 5 for several example trajectories generated by our method in the area1 environment.

SLAM [36] struggles to localize itself using RGBD alone, due to the high complexity of our testing environments. It succeeds only when the start position is close to the goal position. SAC [9] performs much worse than ours due to the sparse reward and limited number of training frames. In [9], each environment is a single room and they use synthetic images but our environment can be up to 1031m² with 40 rooms and we are handling real-world images. Our environment has higher complexity with more obstacles and the entrances to the rooms can be extremely narrow resulting in a difficult solution. SAC needs millions of frames to converge in our environment which is not practical. [9] claims they can generalize to new targets by evaluating only on 10 targets that are several steps away from the training targets. In our experimental setup the targets can be anywhere on the map. A majority of the successes for SAC is when the target location happens to be in a same room as the start location. Our method also requires much fewer simulation steps/training frames (~700K compared to 20M) and less training time (90 GPU hours compared to 300 GPU hours).

Our method with no GPS achieves similar performance to our variant with GPS. As an ablation study, instead of having a separate goal checking model, a done action is generated directly from the policy model, and we find that using a separate goal checking model increases the success rate by 0.2 ~ 0.5. In the cases where the policy model incorrectly identifies done, it either outputs done prematurely or passes the goal without terminating. We intentionally keep the amount of training data roughly the same across all environments to evaluate how the performance changes with the complexity of the environment. When the number of rooms is over 30, our method starts to struggle to get to the goal. Despite the reduction in performance due to environmental complexity, our method performs on average 0.556 and 0.442 better in success rate than SLAM and SAC respectively. Given that we achieve high accuracy on smaller environments, we believe the performance in area1 and area2 will go up if trained on more expert trajectories. A future direction is to analyze the amount of training data needed for a given environmental complexity.

### VI. Conclusions

We proposed a navigation pipeline which does not rely on odometry, map, compass or GPS at runtime and is purely based on the visual input and a novel 8-image panoramic goal. Our method learns from expert trajectories generated using RGBD maps of several real environments. Using robotic simulators with real data and photo-realistic rendering, we are able to efficiently collect a large amount of expert trajectories and train in simulation with real-world data, relieving the need of sim-to-real domain adaptation. Our experiments show that the proposed method 1) achieves better performance than deep RL and SLAM baselines, especially in complex environments with difficult and long-range path solutions, 2) requires fewer training samples and less training time, and 3) can work across different environments given an RGBD map. Our future work will explore using semantic labels as features for learning navigation, and will test in more complex environments.

### TABLE I: Different method results over 5 environments, with the best values in bold. For SPL higher values are better. For OOR lower values are better. Our method with no GPS achieves similar performance to our variant with GPS. Our methods perform best across all environments and metrics.

| Environment     | Model   | Success Rate | SPL   | OOR   |
|-----------------|---------|--------------|-------|-------|
| house2 (66m², 6 rooms) | SLAM    | 0.7575       | 0.5208| 2.726 |
|                 | SAC     | 0.8400       | 0.5620| 3.538 |
|                 | Ours (GPS) | 0.9950       | 0.9810| 1.066 |
|                 | Ours (no GPS) | 0.9875       | 0.9724| 1.053 |
| house1 (89m², 10 rooms) | SLAM    | 0.3575       | 0.2380| 2.809 |
|                 | SAC     | 0.5200       | 0.3648| 2.961 |
|                 | Ours (GPS) | 0.9975       | 0.9811| 1.064 |
|                 | Ours (no GPS) | 0.9225       | 0.8748| 1.252 |
| house17 (220m², 14 rooms) | SLAM  | 0.0900       | 0.0642| 2.344 |
|                 | SAC     | 0.2800       | 0.1281| 5.919 |
|                 | Ours (GPS) | 0.9800       | 0.7853| 2.020 |
|                 | Ours (no GPS) | 0.9150       | 0.7179| 2.389 |
| area2 (1031m², 31 rooms) | SLAM    | 0.0700       | 0.0500| 4.447 |
|                 | SAC     | 0.1700       | 0.1073| 3.997 |
|                 | Ours (GPS) | 0.7250       | 0.5556| 2.504 |
|                 | Ours (no GPS) | 0.6625       | 0.4714| 2.948 |
| area1 (786m², 40 rooms) | SLAM    | 0.0100       | 0.0002| 117.1 |
|                 | SAC     | 0.0425       | 0.0195| 6.967 |
|                 | Ours (GPS) | 0.6600       | 0.3954| 4.504 |
|                 | Ours (no GPS) | 0.5750       | 0.2705| 5.896 |
| area2 (483.4m², 20.2 rooms) | Average | 0.2570       | 0.1746| 25.89 |
|                 | SAC     | 0.3700       | 0.2363| 4.076 |
|                 | Ours (GPS) | 0.8715       | 0.7393| 2.232 |
|                 | Ours (no GPS) | 0.8125       | 0.6614| 2.707 |
