Deep Learning for Embodied Vision Navigation: A Survey
Fengda Zhu, Yi Zhu, Xiaodan Liang and Xiaojun Chang

Abstract—Navigation is one of the fundamental features of an autonomous robot. And the ability of long-term navigation with semantic instruction is a ‘holy grail’ goals of intelligent robots. The development of 3D simulation technology has enabled a large scale of data to simulate the real-world environment. The deep learning in embodied navigation proves its ability to robustly learn various embodied navigation tasks. However, deep learning in embodied navigation is still in its infancy due to the unique challenges faced by the navigation exploration and learning from partial observed visual input. Recently, deep learning in embodied navigation has become even thriving, with numerous methods have been proposed to tackle different challenges in this area. To give a promising direction for future research, in this paper, we present a comprehensive review of embodied navigation tasks and the recent progress in deep learning based methods. It includes two major tasks: target-oriented navigation and the instruction-oriented navigation.

Index Terms—deep learning, embodied navigation, point navigation, object navigation, vision-language navigation.

1 INTRODUCTION

Building a robot to finish tasks autonomously in place of human is a topic been researched for a long time. Autonomous navigation in a real world environment is one of a necessary ability to build such a robot. It has been widely applied in various areas, such as autonomous driving, robotic vacuum cleaner and rescue robot. However, some challenges are going to be tackled in real world navigation. For example, collecting data from 3D environment is expensive; the gap between defined navigation tasks and real-world tasks are large; learning from large-scale of noisy data makes the model unstable; it is hard to understand complex navigation instruction like natural language, etc.

With the rapid development of 3D simulation technology, various of 3D datasets such as SUNCG [1], Matterport [2] and Gibson [3], to provide assets to build an embodied environment. Based on these datasets, researchers use different simulators to render the 3D assets to simulate a 3D embodied environment. The agent that interact with the environment through its physical body within that environment is named as ‘the embodied agent’ [4], [5]. The simulated environment provide RGB (most of which also include depth) information for the agent to recognize the world. Some simulators like MINOS [6] also provide contact force sensor and localization sensor. The simulated environment enables the navigation agent to sample numerous trajectories as training data. Sampling from the simulated environment largely reduces the cost of data collection.

Autonomous navigation is a board field with lots of challenges to be solved. Various tasks have been proposed to study the embodied navigation problem based in these simulated environments. Anderson et al. [7] and Savva et al. [6] propose PointGoal and ObjectGoal navigation. These two target-driven navigation tasks are fundamental and could be implemented in any embodied environment. Later, Anderson et al. [8] propose Vision-and-Language Navigation (VLN) to link natural language to vision and navigation policy.

In recent advances of embodied navigation, deep learning methods prove its ability of learning a robust navigation model from large scale of data. Zhu et al. [9] firstly propose a deep reinforcement learning [10], [11], [12] framework that converges fast and generalizes well to unseen scene in target-driven navigation tasks. Consequently, more methods contribute on target-driven navigation tasks are proposed [13], [4], [15], [16], [17]. Many papers [18], [19], [20], [21], [22] research on linking natural language to navigation. Some works [23], [24], [25], [26], [27] focus on transferring models obtained by simulated environments to realistic environments. Several works [7], [28], [29] propose various metrics to better evaluate the performance of the navigation agent. DeSouza et al. summarize previous works of traditional navigation model based on geometrical and topological approaches. And Kruse et al. discuss about human robot interactions. Few surveys of robotic navigation [30], [31] are also available.

However, our paper focus on deep learning methods that solve navigation problems. Compared with existing literature on robotic surveys, this work contributes from three aspects:

1) To the best of our knowledge, our paper is the first work to comprehensively study the advance of deep learning methods on embodied navigation tasks.
2) This paper summarizes all recently proposed embodied navigation datasets, simulators and tasks and compare their unique insights.
3) This paper overviews most of recent and advanced deep learning based navigation methods and their motivations and contributions.
4) Our paper groups the recent works to give out some promising directions of embodied navigation.

2 NAVIGATION IN EMBODIED ENVIRONMENT

There is rising interest in learning to navigate in an embodied environment. Compared with learning in the real-world, it has
TABLE 1: Comparison of existing embodied datasets (*: the datasets render only a room as scene).

| Dataset               | Year | Scenes | Rooms | Object Categories | RGB  | Depth  | 2D Semantics |
|-----------------------|------|--------|-------|-------------------|------|--------|--------------|
| Stanford Scene* [32]  | 2012 | 130    | 130   | -                 | synthetic | X      | X            |
| SceneNet* [33]        | 2016 | 57     | 57    | 218               | synthetic | X      | ✔            |
| 2D-3D-S* [34]         | 2017 | 270    | 270   | 13                | synthetic | ✔      | ✔            |
| SUNCG [1]             | 2017 | 45,622 | 775,574 | 84             | synthetic | ✔      | ✔            |
| CHALET [35]           | 2018 | 10     | 58    | 150               | synthetic | X      | X            |
| Matterport3D [2]      | 2017 | 90     | 2,056 | 40                | realistic | ✔      | ✔            |
| Gibson [3]            | 2018 | 572    | 8,854 | 84                | realistic | ✔      | ✔            |
| Replica [36]          | 2019 | 18     | 35    | 88                | realistic | ✔      | ✔            |

Fig. 1: The rendering results of each dataset.

Here, we discuss several advantages: 1) sampling data from embodied environment is fast; 2) embodied environment provides diverse scenes to improve generalization; 3) it is convenient for human annotation. In this section, we first introduce the embodied navigation environments, including the datasets, simulators and tasks. Then we discuss two major problems, target-driven navigation and navigation with natural language, and the corresponding methods to solve these problems.

2.1 Embodied Navigation Environments

Here, we discuss the environments used for embodied navigation. Firstly, we summarize the datasets provide 3D assets and the simulators that rendering the assets and provide interfaces for navigation agents to interact with. Secondly, we compare the motivations and insights of recent proposed navigation tasks and discuss the problems remain to be solved.

2.1.1 Embodied Datasets

An embodied dataset contains 3D assets like textures and meshes for rendering and other configuration data like object location, object category and camera pose for high-level tasks. A comparison of the proposed datasets is shown in Tab. 1.

Earlier work focus on rendering synthetic RGB views [32]. It train a probabilistic model to generate synthetic data based on hand-created scenes. Later, SceneNet [33] introduce a generator model to annotate 2D semantics. As depth channel is proved to be helpful for navigation agents [13], [37], 2D-3D-S [34] provide assets with depth information. Different from previous works that render in a room as a scene, SUNCG [1] provides a large number of scenes consist of bedroom, livingroom, bathroom, kitchen, etc. However, these datasets use synthetic views that largely different from the real-world imagery, which limits the application of the datasets. Matterport3D [2] provide photo-realistic panoramic views by 3D reconstruction and 2D and 3D semantics of these views. Gibson [3] provide a more diverse dataset with 572 houses. Replica [36] propose a dataset with 18 indoor scene consist of dense meshes and high-resolution textures. Some work such as AI2-THOR [38], RoboTHOR [39] and CHALET [35] rely on the datasets that not currently released. The rendering results of of some datasets are shown in Fig. 1. Note that SUNCG, CHALET and AI2-THOR provide synthetic views that far from real-world scenes, which limit their application.

2.1.2 Embodied Simulators

An embodied simulator provide an interface for an agent to interact with the environment. We compare different features of the existing simulators in Tab. 2. Simulator equips lots of sensors for the agent, such as RGB sensor, depth sensor, physics sensor and location sensor, etc. All simulators equip RGB sensor, depth sensor and location sensor. Early works provides low RGB resolution due to the limit of 3D rendering technology. The lack of visual details limits the navigation performance of the agent. Simulations such as Matterport3D simulator [8], Gibson simulator [3] and Habitat [40] propose high-resolution photo-realistic panoramic view to simulate more realistic environment. Rendering frame rate is also important to embodied simulators since it is critical to training efficiency. MINOS [6] runs more than 100 frame per second (FPS), which is 10 times faster than its previous works. Habitat [40] runs more than 1000 FPS on 512 × 512 RGB+depth image, making it become the fastest simulator among existing simulators. Discrete state space in [8] simplify the navigation problem and make the agent easy to learn complex vision-language navigation tasks. However, continuous state space is more welcomed since it facilitates transferring a learned agent to a real-world robot. A customizable simulator is able to generate more diverse data by changing the object textures and reconfigure the lights. That makes the navigation agent to learn a robust policy. Some complex tasks may require a robot to interact with objects, such as picking up a cup, moving a chair or opening a door. AI2-THOR [38], iGibson [41] and RoboTHOR [39] provide interactive environments to train such a skill. Multi-agent reinforcement learning [42], [43] is a rising problem of cooperation and competition among agents. AI2-THOR and iGibson also support multi-agent training for collaborative tasks.

2.1.3 Embodied Navigation Tasks

Here, we introduce several tasks that study the embodied navigation problem. These tasks are divided into two classes: target-driven tasks and cross-modal tasks. In target-driven tasks, an agent is require to a specific target. The target could be a position, an object and a room. In cross-modal tasks, the instruction is a natural
language, which indicates a target to navigate to or a trajectory it should follow.

**PointGoal Navigation**, firstly defined by Anderson et al. [7], is a task where an agent is initialized to a random starting position and orientation then asked to navigation to a target position. The target position is indicated by its relative coordinates to the starting position. In this task, no ground-truth map is available and the agent must only use its sensory input to navigate. This task requires the agent to accumulatively estimate the distance and directions of its navigation trajectory. This task does not need any human labels since the starting position and the target position are able to be sampled from the environment. Theoretically, this task is able to be applied on all embodied environments.

**ObjectGoal Navigation** is proposed by Zhu et al. [9]. In this task, an agent is initialized to a random starting position and asked to find a specific object like a table or bed. Once the navigation agent finds the target, it stops. The navigation process is regarded as a success if the agent is located within a distance to the target object. ObjectGoal navigation requires the object labels and locations. This task requires the agent to navigate and explore the house efficiently and accurately recognize the target object.

**RoomGoal Navigation** is proposed by Wu et al. [44]. In this task, an agent is initialized to a random position and asked to navigate to a room (e.g. bedroom or kitchen). The navigation process is regarded as a success if the agent is located within the target room. RoomGoal navigation requires the room labels and regions. The concepts of rooms are high-level semantics. Thus it requires the navigation agent to have higher understanding of the scene based on the visual details such as the types of furniture and the arrangement of the room.

**Multi-Object Navigation (MultiON)** Recently, more and more researchers pay attention to long-term attention. Motivated by this, Wani [45] propose MultiON, a benchmark for Multi-Object Navigation. In MultiON, an agent is asked to navigate to multiple target objects one-by-one, which makes the navigation trajectory quite long. The agent uses a FOUND action to declare that it reaches the instructed target. Perception and effective planning under partial observation would be the key to solve this task. There are two simulators that support multi-agent navigation: the Habitat and the iGibson.

**Vision-and-Language Navigation (VLN)** is a task where an agent navigates step-by-step following a natural language instruction [8]. This task bridges multiple fields, such natural language processing, computer vision, reinforcement learning and robots. VLN is a significant step forward the real-world robot applications since it is the first work that introduce natural language into embodied navigation. Previous tasks such as ObjectGoal and RoomGoal hard-code the object and room semantics as a one-hot vector. On the contrary, VLN introduce natural language sentences to instruct the navigation process like **Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.** The VLN task is successfully completed if the agent stops close to the intended goal following the instruction. To accomplish this task, a robot has to learn to align vision and language modalities and make decisions based on the result of cross-modal alignment.

The room-to-room (R2R) dataset is proposed in [8] to validate the VLN. The R2R dataset contains 21,567 navigation instructions with an average length of 29 words. Three different instructions are used to describe a trajectory, which ensures the language diversity. The average trajectory length is 10 meters. Since the trajectories in R2R dataset are annotated on Matterport3D Simulator, they are represented by a discrete path in a graph. Each trajectory is 6 hops. However, due to the domain limitation of this task, the instruction vocabulary is relatively constrained, consisting of 3.1k words.

However, the R2R dataset has several shortcomings: 1) the referenced paths are direct-to-goal so that R2R instructions lack of capability of describing complex paths; 2) the instruction consists of several sentence and not finegrained; 3) the scale of the training data is small and adding augmented data could improve navigation performance; 4) The instruction language is English only and does not include other languages. To address these problems, lots of VLN datasets are proposed, such as R4R [46], FGR2R [47] and RxR [48]. More detail

**Navigation from Dialog History (NDH)** When navigating in an unfamiliar environment, a human usually ask for assistance and continue navigation according to human responses. However, building an agent that is able to autonomously ask natural language question and react to the answer is still a long-term goal in robotic navigation. Thomason et al. [49] proposes NDH where an agent navigates according to a dialog history consists of several question-answering pairs. They suggest that researching on NDH is fundamental for building a real dialog navigation robot.

**Embodied Question and Answering (EQA)** Visual Question Answering (VQA) [50] is a computer vision task where a system answers a text-based question given an image. VQA soon became one of the most popular computer vision tasks since it reveals a possibility that a human interacts with an AI agent with natural language [51], [52], [53]. Compared with VQA, a more advanced activity is to answer question by self-exploration in an unseen environment. Das et al. [54] propose Embodied Question and Answering (EQA), where an agent is spawned at a random location in a 3D environment and asked a question. EQA is a challenging task since it requires a wide range of AI skills: visual perception, language understanding, target-driven navigation, commonsense reasoning, etc. In addition to navigation accuracy in other tasks,
EQA propose \textit{EQA accuracy} to measure if the agent correctly answers the question or not.

**Multi-Target Embodied Question and Answering (MT-EQA)**

The natural language questions in EQA is simple since each of them describes one object and lack of attributes and relationships between multiple targets. To tackle this challenge, Yu \textit{et al.} [55] propose MT-EQA. In this task, the instructions are like “\textit{Is the dresser in the bedroom bigger than the oven in the kitchen}”, where the \textit{dresser} and the \textit{oven} locate in different places with different attributes. Thus the agent has to navigate to multiple places, find all targets, analysis the relationships between them, and answer the question.

**Interactive Questioning and Answering (IQA)**

Building an agent which is able to interact with a dynamic environment is long-standing goal of AI community. Recently proposed interactive simulators [38], [39], [41] provide basic functions like open a door or move a chair, which enables researchers to build an interactive navigation agent. Gordon \textit{et al.} [56] propose Interactive Questioning and Answering (IQA), a task that requires an agent to answer questions by interacting with objects in an environment. IQA contains 76,800 training questions that includes existence questions, counting questions, spatial relationship questions.

\textit{“Help, Anna!”} (HANNA) Nguyen \textit{et al.} [57] propose HANNA, an object-finding tasks which allows an agent to request help from Automatic Natural Navigation Assistants (ANNA) when it gets lost. Different from NDH that provides a global dialog history as the instruction, the HANNA offers a environment where the instructions dynamically change by the situation. The environment creates a interface that enables a human to help the agent when it gets lost in testing time.

**REVERIE**

Recently, Qi \textit{et al.} [58] proposes Remote Embodied Visual referring Expression in Real Indoor Environments, named REVERIE in short, to research associating natural language instructions and the visual semantics. Different from VLN that gives an instruction that describe the trajectory step-by-step toward the target, the natural language instruction in REVERIE refers to a remote target object. Thus this task more clearly reflects the capability of natural language understanding, visual navigation, and object grounding of a navigation agent. Compared with ObjectGoal navigation, REVERIE offer rich language descriptions to enable the agent to find a unique target in the house. The REVERIE dataset is built upon Matterport3D Simulator, which comprises 10,567 panoramas containing 4,140 target objects. The REVERIE dataset annotates 21,702 crowd-sourced instructions with an average length of 18 words and a vocabulary of over 1,600 words. There are 4,140 target objects of 489 categories in the REVERIE, whose categories are 6 times more than the 80 categories in ReferCOCO [59].

**Audio-visual Navigation**

proposed by Chen \textit{et al.} [60] introduces audio modality for embodied navigation environment. This task requires the agent to navigate to a sound object by seeing and hearing. It encourages researchers to study the role of audio plays in navigation. Chen \textit{et al.} also offer the SoundSpaces [60] dataset for the Audio-visual Navigation task. The SoundSpaces dataset is built upon two simulators, Replica and Matterport3D. It contains 102 natural sounds across a wide variety of categories: bell, door opening, music, people speaking, telephone, etc.

### 2.2 Evaluation Metrics

Many evaluation metrics have been proposed to evaluate how well a navigation agent performs. Earlier work [9] use the average number of steps (i.e., average trajectory length) it takes to reach a target from a random starting point. However, there are a large proportion of trajectories fail when the navigation environment become more complex and the navigation task become more challenging. Later works [6], [14], [44] introduce propose the \textit{Success Rate (SR)} to measure frequency of the agent successfully reach the goal and other works [14], [54] report Navigation Error (NE), the mean distance toward the goal when the agent finally stops. success rate in the presence of an oracle stopping rule Oracle Success Rate (OSR) is proposed to evaluate if the agent correctly stops following the oracle stopping rule [8], [61]. These metrics either measures the probability of the agent completes the task or measures how much proportion it completes the task. Anderson \textit{et al.} [7] analysis the advantages and disadvantages of the previous metrics and propose a novel metric, named \textit{Success weighted by Path Length}.

This metric takes both the trajectory length and the success rate into consider. SPL is the first metric that evaluate both the efficiency and efficacy of a navigation agent, and it is regarded as the primary metric in VLN. SPL ignores the turning actions and the agent heading. \textit{Success weighted by edit distance} (SED) [62] takes turning actions into consideration and fix this problem. And SED is designed for instruction following in graph-based environments, where a specific correct path exists.

However, in some tasks like R4R [46] and R6R [63], the instructed paths are not direct-to-goal. Thus, it is inappropriate to evaluate the navigation performance by SPL. Therefore, \textit{Coverage weighted by Length Score} (CLS) [46] is proposed to measure the fidelity of the agent’s behavior to the described path. CLS is the product of two variables: the path coverage and the length score. Ilharco \textit{et al.} absorb the idea of Dynamic Time Warping [64], an approach widely used in various areas [65], [66], [67], and propose normalized Dynamic Time Warping (nDTW) metric [68] to evaluate the navigation performance. Similar to CLS, nDTW evaluates the distance between predicted path with the ground-truth path. Moreover, nDTW is sensitive to the order of the navigation path while CLS is order-invariant. nDTW can be implemented in a efficient dynamic programming algorithm. The path sensitive metrics, like CLS and nDTW, performs better when they are used as reward function than target-oriented reward function in reinforcement learning to navigate [46], [68]. Most of recent works adopt multiple metrics to comprehensively evaluate their proposed models. We compare the existing metrics in Tab. 3. Suppose we have a predicted trajectory \( P \) and a ground truth trajectory \( R \). \( p_i \) and \( r_i \) are the ith node on trajectory \( P \) and \( R \). \(|P|\) and \(|R|\) stand for the length of \( P \) and \( R \) respectively. We measure the score of the agent from the below aspects:

- **Path Similarity (PS)** characterize a notion of similarity between the \( P \) and the \( R \). This implies that metrics should depend on all nodes in \( P \) and all nodes in \( R \). PS penalizes deviations from the ground truth path, even if they lead to the same goal. This is not only prudent, as agents might wander around undesired terrain if this is not enforced, but also explicitly gauges the fidelity of the predictions with respect to the provided language instructions.

- **Soft Penalties (SP)** penalizes differences from the ground truth path according to a soft notion of dissimilarity that depends on distances in the graph. This ensures that larger discrepancies are penalized more severely than smaller ones and that SP should not only rely on dichotomous views of intersection.

- **Unique Optimum (UO)** yields a perfect score if and only if the reference and predicted paths are an exact match. This ensures that the perfect score is unambiguous: the reference path \( R \) is therefore...
The model-free methods learn to navigate end-to-end without modeling the environment. The learning objective include imitation learning or reinforcement learning. Though extensive reinforcement learning works [61], [69], [70] have long studied 2D navigation problem where an agent receives global state for each step, the embodied navigation problem with partial observation remains challenging. Many robot control works [71], [72], [73], [74] focus on obstacle avoidance rather than trajectory planning.

Zhu et al. [9] firstly propose to use deep learning for feature matching and deep reinforcement learning for policy prediction. The proposed framework allows the agent to better generalize. And they propose to train navigation model in a simulated environment for the sampling efficiency of data. Sequentially, Successor Representation (SR) [75] is proposed to enable the agent to interact with objects. This framework takes the states of objects and a discrete description of the scene into consideration. It encodes these semantic information and concatenate with the visual representation as in [9]. Different from [9] that uses reinforcement learning only to learn a policy predictor, Successor Representation model bootstraps reinforcement learning with imitation learning. Wu et al. [44] builds an agent that is able to generalize to unseen scenes. Since the navigation agent receives a partial observation as the input, this work advocates to introduce an LSTM layer to encode historical information. By ablating different RL algorithms, this work proves that A3C [76] outperforms DDPG [77] in navigation task. The model is built upon House3D simulator [44] that provides 45,622 human designed scene with segmentation annotation. The model learned from semantic mask outperforms which learned from RGB inputs. It indicates that semantic information is much more important than pure visual inputs. However, the visual images in House3D are rendered by computer graphic, which are largely different from the observations from the real-world environment. Li et al. [78] propose a end-to-end model based on Q-learning that learns viewpoint invariant and target invariant visual servoing for local mobile robot navigation.

There are lots of work use the segmentation to augment visual inputs. Mousavian et al. [79] introduce a Faster-RCNN detector trained on MSCOCO dataset [80] and a segmenter defined by [81] to detect and segment objects and fed the features into the policy network. Shen et al. [82] propose a framework that fuse diverse visual representations, including RGB features, depth features, segmentation features, detection features, etc. The different visual representations are adaptively weighted in fusing. To further improve the robustness, they propose a inter-task.
affinity regularization that encourages the agent to select more complementary and less redundant representations to fuse. Ye et al. [83] propose a hierarchical two-layer structure, where the high-level layer plans over the sub-goal, and the low-level layer plans over the atomic actions to achieve the goal. Despite the well-performed detector and segmenter, it is still hard for the agent to learn navigation from the environment. When a human learns to navigate, he/she recall the prior knowledge that he/she learns in his/her past experience. For example, to search for mugs, a human would search cabinets near the coffee machine and for fruits a human may try the fridge first. Lv et al. [84] integrate 3D knowledge graph and sub-targets into deep reinforcement learning framework. They further use attention mechanism that learns to focus on the most important references in the current field of view or target image to guide policy search. Another work [85] proposes a model to learn prior knowledge by a Bayesian relational graph. Then the model estimating posterior layout at test time by means of memory update. Therefore, modeling the attributes and relationships of semantic concepts as a graph is proven to be beneficial for navigation. To enhance the cross-target and cross-scene generalization, Wu et al. [86] introduce an information theoretic regularization term into the RL objective. In this way, the agent models the action-observation dynamics by learning a variational generative model and improve generalization in cross-target and cross-scene tasks.

Data efficiency is a long concerned problem in training a navigation agent. Lots of works have proven that larger batch and longer training period significantly improve the navigation performance. However, to achieve that optimal performance is impractical in most of experiments. Based previous works on synchronous distributed RL [87], [88], Wijmans et al. [89] present Decentralized Distributed Proximal Policy Optimization (DD-PPO), a distributed reinforcement learning methods for efficient training. DD-PPO exhibits near-linear scaling, which achieves a speedup of 107x on 128 GPUs over a serial implementation. With 64 GPUs, DD-PPO is able to accomplish 2.5 Billion steps of experience (over 6 months of GPU-time training) in under 3 days of wall-clock time.

Some works investigate problem settings other than indoor navigation, such as street view navigation or combining other modalities. Khosla et al. [90] firstly reveals the possibility that an agent navigation based on the visual cues of street views. Consequently, Brahmbhatt et al. [91] present DeepNav, a Convolutional Neural Network (CNN) based algorithm for navigating large cities using locally visible street-view images. These works rely on supervised training with the ground truth compass while the compass can be unavailable in real-world. Mirowski et al. [92] build an end-to-end deep reinforcement learning approach that relies on the street scene as visual input only without the ground truth compass. Recognizing the importance of locale-specific knowledge to navigation, they propose a dual pathway architecture that allows locale-specific features to be encapsulated. Another contribution of this work is that it proposes an interactive navigation environment that uses Google Street View for its photographic content and worldwide coverage. Gan et al. propose a task where an agent receive audio signal in addition to visual images to navigate. This work demonstrates that existing state-of-the-art deep reinforcement learning approaches meet difficulties in generalization. To tackle this challenge, Chen et al. [93] proposes AV-WaN, a model that learn to set audio-visual waypoints. The agent dynamically sets intermediate goal locations based on its audio-visual observations and partial map in an end-to-end manner.

### 2.3.2 Self-Supervised Methods

Self-supervised learning is a long studied topic of exploiting extra training signals via various pretext tasks. It enables an agent to learn more knowledge without additional human annotations. Various self-supervised tasks have been proposed in the field of deep learning, such as context prediction [94], solving jigsaw puzzles [95], colorization [96], rotation [97]. There is also some auxiliary tasks proposed to improve data efficiency and generalization in reinforcement learning. Xie et al. [98] combines self-supervised learning with model-based reinforcement learning to solve robotic tasks. Motivated by traditional UVFA architecture [99] which learns a value function by means of feature learning, Jaderberg et al. [100] invent auxiliary control and reward prediction tasks that dramatically improve both data efficiency and robustness.

In embodied navigation, the environment contains unstructured semantic information that is hard to learn in end-to-end manner. In spite of explicitly modeling the environment using SLAM or memory mechanism, self-supervised learning provide another feasible way of learning the unstructured knowledge. Mirowski et al. [13] propose an online navigation model with two auxiliary learning objectives, predicting the current depth view by RGB view and detecting the loop closure. Dosovitskiy et al. [101] present a model play the first-person game Doom [102]. In this game, actions include navigating and shooting. This model learn to act based on raw sensory input from this complex three-dimensional environment. Visual perception is critical for visual navigation. But the training signal provided by reinforcement learning contain too much noise to train a robust feature perception network. Ye et al. [103] introduce an encoder-decoder architecture that encodes the current visual view and predicts its segmentation. An auxiliary task is used to penalize the segmentation error, which encourage the learning of feature perception. However, these auxiliary tasks only exploit the simple dynamics of the between two adjacent
states. Liu et al. [104] propose an auxiliary regularization task to guarantee the consistency of actions in a trajectory. The auxiliary task penalizes the inconsistency of representations to encourage the policy network to extract salient features from each sensor. Real-world robot locomotion is far from deterministic due to the presence of actuation noise, which might be caused by wheels slipping, motion sensor error, rebound, etc. To reduce the noise, Datta et al. [105] introduce an auxiliary task of localization estimation based on the idea of temporal difference. The auxiliary task is used to train a CNN network and use the estimated locomotion as an input of the policy network. This work is the runner-up in the PointNav track of CVPR 2020 Habitat Challenge. Bigazzi et al. [106] introduce a curiosity-driven self-supervised objective to encourage exploration while penalizing the repeating actions. A stable curiosity-driven policy without repeating actions improves the exploration efficiency. Dean et al. [107] use audio as an additional modality for self-supervised exploration. It includes an curiosity driven intrinsic reward, which encourages actions to generate novel associations between different sensory modalities (audio and visual). An overview of the pipeline of self-supervised navigation methods is shown in Fig. 2.

### 2.3.3 Planning-based Methods

The map building problem for an unknown environment while solving the localization problem at the same time is known as Simultaneous Localization and Mapping (SLAM) [108], [109]. The earlier investigations on visual navigation were carried out with a stereo camera [110], [111] and a monocular camera, such as MonoSLAM [112]. Over the past decade, traditional geometric-based approaches [113], [114], [115] still dominate the field. With the development of deep learning, some methods like CNN-SLAM [116], DVO [117] and D3VO [118] are proposed. Some indoor tasks are proposed to study SLAM, such as KITTI [119] and EuRoC [120]. However, these tasks are different from embodied navigation task. The odometry benchmark is to estimate the location given a sequence visual inputs while the navigation task is to align the instruction with the environment semantics.

Recently, researchers discover that the ability of localization is important to reinforcement learning and navigation. Thus, some work have been proposed to introduce SLAM methods to help navigation, especially the long-term trajectory planning. Gupta et al. first propose to use SLAM-based method in navigation [121]. This work consists of two parts: mapping and planning. The mapping mechanism maintains a 2D memory map. For each step, it transforms the embodied scene into a 2D feature and update the map with the feature. The planning mechanism uses a value function to output a policy. Neural Map [122] generalize this idea for all deep reinforcement learning agents rather than navigation only. However, this work does not utilize the 2D structure of this memory as all their operations can be conducted and assume the location of the agent is always known. Neural SLAM [123] fix this problem by embedding SLAM-like procedures into the soft-attention [124]. To avoid spatial blurring associated with repeated warping, MapNet [125] proposes to use a world-centric rather than an egocentric map. Different from previous works, MapNet maintains a 2.5D representation by a deep neural network module that learns to distill from RGB-D input. Gordon et al. [56] proposes Hierarchical Interactive Memory Network (HIMN), a framework with hierarchical controller for IQA task. The high-level controller is a planner that decides the long-term navigation target and the low-level controller predicts the action, interacts with the environment, and answers the question. The real-world navigation is much complex than which in simulated environments. Qi et al. [126] advocated that autonomous agents must follow a navigation policy that avoids collisions with dynamic obstacles to ensure safe operation. Motivated by this, their work takes the affordance and semantic constraints into consideration. This work learns a spatial affordance map to instruct the agent where the it can safely move. Georgakis et al. [127] proposes to encode the spatial map by two convolutional layers following an LSTM layer since the spatial map during navigation is a partial observation of the total scene and the LSTM layer is able to maintain historical information. Yang et al. advocate in [128] that to integrate semantic and functional priors is a key to navigation and build a topological semantic graph in the context of encoding semantic scene priors. To obtain a better understanding of visual features, Yang et al. adopt a Graph Convolutional Network (GCN) [129]. More generally, Chen et al. [130] propose to use GNNs to pose robot navigation as a graph traversal problem in a topological map of the environment. Narasimhan et al. [131] propose a model to learn the top down belief maps containing room semantics. Different from classical SLAM approaches that learns the house layout without classifying the rooms, this method learns the beliefs for each type of room respectively.

Efficient exploration is widely regarded as one of the main challenges in reinforcement learning (RL) [132], [133], [134], [135], [136], [137], [138]. Similarly, it is important in navigation since the target does not always visible from the starting position and the agent is required to explore the unseen scene and search for the target. Recently, exploration based on explicitly modeled semantic memory is proven to be efficient. Chen et al. [139] find that use of policies with spatial memory that are bootstrapped with imitation learning and finally finetuned with coverage rewards derived purely from on-board sensors can be effective at exploring novel environments. Active Neural SLAM (ANS) [16] is a success neural SLAM method which achieves the state-of-the-art on the CVPR 2019 Habitat Pointgoal Navigation Challenge. ANS proposes a hierarchical structure for planning. Based on the idea of hierarchical RL [140], [141], [142], [143], ANS learns the high-level planner by reinforcement learning and learns the low-level planner by imitation learning. The mapper is implemented by an auxiliary task of predicting a 2D map. The first channel of the map stands for if there is an obstacle and the second map stands for if the position has been explored. However, the predefined 2D map is not helpful for long-term navigation. Thus a more flexible way is proposed to store the observed feature named Neural Topological SLAM [144]. This method introduce a graph update module to leverage semantics. The graph update module maintains a topological feature memory. For each step, the module localize current observation into memory nodes. If an observation is not localized in any node of the memory, the graph update module will add a new node into the topological feature memory. Goal-Oriented Semantic Exploration (SemExp) [17] tackles the object goal navigation task in realistic environments. This method first builds a episodic semantic map and uses it to explore the environment based on the goal object category. The methods achieves state-of-the-art in Habitat ObjectNav Challenge 2020. An overview of the common practice of the ‘Neural SLAM’-based model is shown in Fig. 3.
2.4 Summary

Compared with traditional robotics methods, the model-free method are able to obtain robust navigation models by sampling large scale of data with the embodied simulator. Some works adopt detection and segmentation approaches to get better perceive visual views. In spite of indoor scenarios, model-free methods achieve great success in street scene and multi-modal environments. Compared with other visual tasks like VQA or image captioning, visual navigation task provide an embodied environment which contains rich semantic information other than visual inputs. Motivated by this, self-supervised methods are proposed to exploit the extra knowledge by auxiliary tasks to improve the learning efficiency and generalization. However, these methods lack of exploration efficiency since the methods do not consider to explicitly modell the environment. Planning-based methods learn to dynamically model the environment during navigation. These methods use a 2D map or topological memory to store the features of seen visual views. By modeling the environment, the agent is able to avoid some repeat actions and navigate more efficiently.

2.5 Navigation with Natural Language

A navigation robot which is able to understand natural language can accomplish more complex tasks, such as “pick up the cup in the kitchen” or “help me find my glass upstairs”. Recently, navigation with a natural language instruction is attracting rising attention due to its wide range of application. Researchers have proposed diverse vision-language navigation tasks. Each of task has its own challenges. In this section, we introduce three kinds of works that solves these challenges: 1) end-to-end methods; 2) pretraining-based methods; 3) planning based methods.

2.5.1 Step-by-step Navigation

Anderson et al. [8] proposes the vision-language navigation task, where an agent navigate to a target following a natural language instruction. This work provide the Room-to-room dataset containing more than 21K instructions with 29 words on average. Each language instruction correspond to a navigation trajectory. It is the first task that introduce natural language into visual navigation. Consequently, lots of vision-language navigation methods have been proposed to research on building an agent that perceive both visual and language information.

Anderson et al. propose a sequence-to-sequence model in [154] to solve the vision language navigation problem. This model sequentially encodes a language instruction word-by-word, concatenates the sentence feature with the visual image feature and decodes the action sequence. However, this model is lack of stability and generalization since it fails to consider the dynamics in the real-world environments. Wang et al. [145] propose a model-based method to bridge the gap between synthetic studies and real-world practices. This work introduces an environmental module that learns to predict the future state and reward. This module helps improve the generalization of unseen scenes. Even though the data scale provided by [8] enables researchers to train a stable navigation model, it is lack of both visual and language diversity and the trained model is easily overfits to the seen scenes and perform badly on the unseen scenes. Motivated by this, Fried et al. [18] propose a data augmentation approach to improve the model generalization. This work proposes a speaker-follower model which consists of two submodules: the speaker which translate a trajectory into an instruction and the follower which translate an instruction into a trajectory. Fried et al. generate augmented data by randomly sampling trajectories in the environment and use the speaker to translate the trajectories into instructions. In addition, Fried et al. define a high-level action in [18] that move forward toward an orientation in a panoramic space in stead of low-level actions like turn left, turn right and go forward. Compared with the definition of low-level actions, this approach largely reduce the length of the action sequence that describes the same trajectory. Therefore, it needs less prediction times and make the model easier to train and more robust in testing. Previous methods learn to navigate by imitation learning with the instruction-trajectory data pairs. However, imitation learning only supervise the shortest path while ignore the sub-optimal trajectories, which makes the model easily overfits to the training data. To tackle this problem, Wang et al. [19] propose a method combining imitation learning with reinforcement learning. In addition, this method introduce an LSTM to encode the temporal information of visual features and introduce a cross-modal mechanism to achieve better vision-language navigation ability. Ma et al. [20] propose a self-monitoring agent with a visual-textual co-grounding module and progress monitor. The progress monitor use the cross-modal feature from the co-grounding module and estimate the completed progress. Since the instruction in vision-language task guides the agent to the target step-by-step, the progress information contain rich knowledge that help improve the perception of the agent. Moreover, the progress monitor also helps navigation planning. Ma et al. propose in [21] the Regretful Agent, with a regretful module which uses the estimated progress to indicate if the agent navigates to a wrong place and need to go back. Similar to the Regretful Agent, Ke et al. propose [147] a framework for using asynchronous search to boost a VLN navigator by enabling explicit backtrack. Anderson et al. [155] regard the step-by-step navigation process as a visual tracking task. This implements the navigation agent within the framework of Bayesian state tracking [156] and formulates an end-to-end differentiable histogram filter [157] with learnable observation and motion models. Since the matterport3D dataset only provide 61 scenes for training, which lacks of visual diversity, the learned model is easily overfits to the training scene. One commonly used method that relieve the visual overfitting is to apply a dropout [158] layer on the visual feature, which is extracted by a pretrained network like VGG [159] or ResNet [160]. Tan et al. [148] argue that simply applying a dropout layer on the visual feature lead to inconsistency, e.g. a chair in this frame could be dropped in the next frame. To solve the problem, Tan et al. propose a environmental dropout layer in [148] that dropout some fixed channels during a trajectory. Different channel sampling for the same instruction-trajectory pair is regarded as a sort of data augmentation. By means of that, the model obtain better visual generalization. Zhu et al. [22] propose AuxRN, a framework that introduce self-supervised auxiliary tasks into vision-language navigation to exploit environmental knowledge. In addition to the temporal difference auxiliary task which is widely use in other vision navigation methods, Zhu et al. propose a trajectory retelling task and instruction-trajectory matching task that learn the temporal semantics of a trajectory. AuxRN [22] improves the progress matching task. Since the standard VLN benchmark is tested based on discrete action space, Zhu et al. discover that supervising the progress with normalized hop number is better than the normalized distance. Instead of generating the low-quality augmented data, Fu et al. [161] introduce the concept of counterfactual thinking to sample challenging paths that force the navigator to improve. They
present a model-agnostic adversarial path sampler (APS) to pick the difficult trajectories and only consider useful counterfactual conditions.

Different from the earlier works that based on data augmentation and other classical navigation methods, some works discover the importance of natural language to VLN. Thomason et al. [162] find the unimodal baseline outperforms random baselines and even some of their multimodal counterparts. Thus the work advocates that ablating unimodal to evaluate the bias is important to proposing a dataset. A study of Huang et al. [163] shows that only a limited number of those augmented paths in [18] are useful and after using 60% of the augmented data, the improvement diminishes with additional augmented data. This work discover that the VLN agent warm-started with pre-trained components from a cross-modal discriminator outperforms the augmentation approaches without any additional human annotations. Reinforcement Learning (RL) make VLN models better generalized by encouraging sub-optimal solutions. To avoid the extensive work in reward engineering, Wang et al. [164] propose a Soft Expert Reward Learning model that includes two parts: 1) soft expert distillation, encourages agents to behave like an expert in soft fashion; 2) self-perceiving, which pushes the agent towards the final destination as fast as possible. Xia et al. [165] leverages multiple instructions as different views for the same trajectory to resolve language ambiguity and improve generalization. It indicates that the human annotations in VLN are largely biased according to the specific scene and the trajectory. Previous works extract global visual features from panoramic views by a pretrained CNN network like ResNet-101. Hong et al. [152] introduce Faster-RCNN to detect objects in navigation and build a relationship graph between visual and language entities for vision-language alignment. Hu et al. [166] discover that the models which only use route structure instead of visual features, outperform their visual counterparts in unseen new environments on the VLN benchmark. They conclude that both the language and the route structures contains high-level semantic information while pixel-based visual representations are a lower-level modality, which makes the vision-language alignment difficult. Motivated by this, Hu et al. propose a novel framework in [166] that decomposes the grounding procedure into a set of expert models with access to different modalities and ensemble them at prediction time. To better research what role language understanding plays in VLN task, Jain et al. [46] point out that the previous metric, such as NE, SR and SPL, depend only on goal instead of the actual path. And current training data in the R2R dataset are direct-to-goal. Therefore, they propose a new metric, Coverage weighted by Length Score (CLS), that measures how much the agent navigates to cover the described trajectory. Hong et al. [47] argue that the intermediate supervision is important in vision-language alignment. Thus, they propose FGR2R, a framework that learns at the level of sub-instructions. Despite of just focusing on the target, this method makes navigation process traceable and encourage the agent to run precisely on the described path. Kurita et al. [167] argues that the rich semantics in natural language is able to guide navigation. Different from previous methods that use a discriminative model to predict actions, Kurita et al. learns a generative model and use Bayes’ rule to build a language-grounded policy. Given the instruction, the model predict the next action by applying Bayes’ rule to obtain the posterior distribution over all actions.

### 2.5.2 Pretraining-based Methods

With further research on the vision-language navigation, several challenges are addressed, include: 1) low training efficiency; 2) large data bias (include both vision and natural language); 3) lack of generalization from seen to unseen scenes.

**Low training efficiency** The traditional encoder-decoder framework first samples the total trajectory by teacher-forcing or student-forcing and then back-propagate the gradients. In other deep learning tasks like image classification or text prediction, the model predicts a result directly. However, in the vision-language navigation task, the agent predicts a trajectory by interacting with the environment step-by-step, which is so time-consuming that reduce the training efficiency.

**Large data bias** The vision-language navigation scenarios are so diverse that 61 houses in R2R cannot cover all of them. From the aspect of natural language, in the R2R task, only 69% of bigrams are shared between training and evaluation.
**Lack of generalization** Even though the trajectory augmentation, visual feature augmentation and natural language instruction augmentation methods have been proposed to reduce the data bias, lack of diverse training data still largely limits the generalization. Thus, introducing extra knowledge from other tasks and datasets becomes a promising way.

Pretraining-based methods largely improve model generalization by learning in large scale of data [51], [160], [168]. Furthermore, bert-based methods [169], [170], [171] pretrains a transformer network with proxy tasks and achieve great success in vision, language and cross-modal tasks. Many researchers consider to solve the vision-language navigation problem by pretraining-based methods. Li et al. [146] propose PRESS first introduce a pretrained language models to learn instruction representations. And they propose a stochastic sampling scheme to reduce the gap between the expert actions in training and the sampled actions in testing. Majumdar et al. [172] advocate to improve model by leveraging large-scale of web data. However, it is hard to transfer the static image data to VLN task. Therefore, they propose VLNbert, a transformer-based model which is pretrained by static images and its caption. Hao et al. [150] suggest that the data sampled from embodied environment also beneficial in pretraining. They propose PREVALENT, a model self-supervised learn from large amount of image-text-action triplets sampled from an embodied environment. PREVALENT is proven to be effective on several vision-language navigation datasets, including R2R, CVDN and HANNA. The embodied navigation agent receives partial observation rather than global observation, which is better to be modeled as a partially observable Markov decision process. Different from the encoder-decoder model, previous pretraining-based models do not memorize previously seen scenes during navigation and utilize temporal knowledge, which causes information loss in action prediction. Motivated by this, Hong et al. [153] propose a recurrent multi-layer transformer network that is time-aware for use in VLN. This method maintains a feature vector to represent temporal context and feed to the network as an input for each step.

### 2.5.3 Navigation with Question and Answering

In stead of passively perceive natural language instructions from a human commander, Das et al. [55] suggest that an intelligent agent should be able to answer a question via navigation. Thus Das et al. present a new task named EQA (Embodied Questioning and Answering), where an agent is spawned at a random location in a 3D environment and asked to answer a question. In order to answer, the agent have to first navigate to explore the environment, gather information through egocentric vision, and then answer the question. To solve this challenging task, Das et al. present PACMAN, a CNN-RNN model with Adaptive Computation Time (ACT) module [173] to decide how many times to repeatedly execute an action [55]. The PACMAN is bootstrapped by shortest path demonstrations and then fine-tuned with RL. However, this method is lack of the ability of high-level representation. In a later work [174], Das et al. propose a hierarchical policy named Neural Modular Controller (NMC) that operates at multiple timescales, where the higher-level master policy proposes subgoals to be executed by low-level sub-policies. Anand et al. [175] discover that a blindfold (question-only) baseline on EQA and find that the baseline perform previous state-of-the-art models. They suggest that previous EQA models are ineffective at leveraging the context from the environment and the EQA1 dataset has lots of noise. Wu et al. [176] propose a simple supervised learning baseline which is competitive to the state-of-the-art EQA methods. To improve EQA performance in unseen environment, Wu et al. propose a setting in which allows the agent to answer questions for adaptation. Yu et al. [55] argues that the EQA task assumes that each question has exactly one target, which limits its application. Therefore, Yu et al. present Multi-Target EQA (MT-EQA), a generalized version of EQA. The question of this task contains multiple targets. And it require the agent to perform comparative reasoning over multiple targets rather than simply perceive the attributes of one target. Wijmans et al. [177] extend the EQA problem to photorealstic environment. In this environment, Wijmans et al. discover that point cloud representations are more effective for navigation. Luo et al. [178] suggest that the visual perception ability limits the performance of the EQA. They introduce Flownet2 [179], a high-speed video segmentation framework as a backbone to assist navigation and question answering. Li et al. [180] propose a MIND module that model the environment imagery and generate mental images that are treated as short-term subgoals. Tan et al. [181] investigate the questioning and answering problems between multiple targets. In this task, the agent has to navigate to multiple places, find all targets, answer the relationships between them, and answer the question. Motivated by many works in Visual Question Answering (VQA) [50], [51], [182], [183] and Video Question Answering (VideoQA) [184], [185], [186], Cangea et al. [187] propose VideoNavQA, a dataset that contains pair of questions and videos generated in the House3D environment. This dataset builds the gap between the VQA and the EQA. This task represents an alternative view of the EQA paradigm: the navigation aspect is made trivial by providing nearly-optimal trajectories to the agent. Deng et al. [188] propose Manipulation Question Answering (MQA) where where the robot is required to find the answer to the question by actively exploring the environment via manipulation. To suggest a promising direction of solving this problem, they provide a framework which consists of a QA model and a manipulation model. Nilsson et al. [189] research on embodied visual active learning, a setting where an agent in a 3D environment explores and occasionally requests annotation.
Paper claims four challenges in modeling turn-based dialogues, which includes: 1) deciding when to ask a question; 2) Generating navigator questions.; 3) Generating guide question answers; 4) Generating navigator actions. In this paper, Roman et al. introduce a two-agent paradigm, where one agent navigates and asks questions while the other guides agent answers. Different from previous works that guide navigator with template language, this work initialize the oracle model via pretraining on CVDN dialogues to generate natural language. A dialog does not always describe a step-by-step navigation process. Rather, the oracle describes the target scene and let the navigator to find it, which commonly occurs when someone get lost in a new building. Hahn et al. [198] propose a LED task (localizing the Observer from dialog history) to realize such a scenario. Based on this scenario, they present a dataset named WHERE ARE YOU [198] that consists of 6k dialogs of two humans. Due to the wide application of multi-agent communication systems [199], [200], [201], [202] in real-world, researchers become interested in implementing dialog navigating in physical environments. Marge et al. [203] present MRDwH, a platform that implements autonomous dialogue management and navigation of two simulated robots in a large, outdoor simulated environment. These benchmarks are all limited to simulations, while deployed robots operate in noisier, real-world environments. Banerjee et al. [204] propose RobotSlang benchmark, by contrast, was gathered by pairing a human “driver” controlling a physical robot and asking questions of a human “commander”, where the pair need to perform cooperative global localization while carrying out a navigation task to multiple object targets.

We compare the difference of Embodied Question Answering (EQA) [54], Multi-Target Embodied Question Answering (MT-EQA) [55] and Vision-and-dialog navigation (VDN) [49] in Fig. 6. We demonstrate three different dialogues for the same navigation task does not consider the interaction between the agent and the instructor, which always happen when the agent get lost during navigation. In the field of embodied navigation, Banerjee et al. [57] propose “Help, Anna!” (HANNA), an interactive photo-realistic simulator in which an agent fulfills object-finding tasks by requesting and interpreting natural language and vision assistance. Nguyen et al. [196] propose a task named VLNA, where an agent is guided via language to find objects. However, the language instruction in these two tasks far from real-world problem: the responces of HANNA are automatic generated from a trained model while the guidance of VLNA are in the form of templated language that encodes gold-standard planner action. Vries et al. [197] propose “Talk The Walk” (TTW), where two humans communicate to reach a goal location in an outdoor environment. However, in TTW, the guiding human uses an abstracted semantic map rather than an egocentric view of the environment. Thomsan et al. [49] propose vision-and-dialog navigation (VDN), a scaffold for navigation-centered question asking and question answering tasks where an agent navigates following a multi-round dialog history rather than an instruction. Each round coorespond to a sub-trajectory. The more fine-grained annotation facilitate researchers to study the problem of navigation with natural language language. The dialogues contain complex phenomena that require egocentric visual grounding and referring to both dialog history and past navigation history for context. Zhu et al. [149] propose a framework with a cross-modal memory mechanism to capture the hierarchical correlation between the dialogue rounds and the sub-trajectories. More generally, several methods, such as PREVALENT [150] and BABYWALK [63], validate their navigation ability for both sentence instructions and dialog instructions. Different from previous works, Roman et al. [190] focus on asking questions rather than understanding answers.
2.6 Summary
Natural language provides an interface for a human to interact with a robot. An robot equipped with a skill of natural language understanding is able to accomplish a lot of complex tasks such as navigating following a natural language instruction or a dialogue, asking the oracle for more details, etc. Lots of works have been proposed to research on vision-language navigation problem from diverse aspects. In the future, researchers will focus on complex vision-language scenarios and enable the robot to work on more real-world like settings.

3 Real-world Embodied Navigation
Many real-world navigation applications achieve great success due to the development of deep learning, reinforcement learning and planning, etc. Moreover, embodied navigation methods that trained in a simulated environment give a promising direction for real-world navigation applications or improve the performance of these real-world applications by sim-real transferring methods. In this section, we are going to 1) introduce some embodied navigation methods for different real-world applications; 2) compare them with the methods in simulators; 3) discuss the possibility of sim-to-real transferring.

3.1 Methods for Real-world Navigation
Deep learning plays an important role in indoor navigation for real-world applications. LeCun et al. [205] firstly adopt convolutional network for obstacle avoidance. Hadsell et al. [206] propose a self-supervised learning process that accurately classifies long-range vision semantics via a hierarchical deep model. The method is validated on a Learning applied to ground robots (LAGR) [207] mobile robotic vehicle. Later, more and more works adopt deep learning to perceive and extract distinctive visual features [208], [209], [210]. Zhang et al. [15] research on the problem where a real robot navigates in simple maze-like environments. Based on the success of RL algorithms for solving challenging control tasks [10], [77], [211], [212], Zhang et al. employ successor representation in learning to achieve quick adaptation. Morad et al. [213] present an indoor object-driven navigation method named NavACL that uses automatic curriculum learning and is easily generalized to new environments and targets. Kahn et al. [214] adopt multitask learning and off-policy RL learning to learn directly from real-world events. This method enables a robot be able to learn autonomously and be easily deployed on multiple real-world tasks without any human provided labels.

There has been a long history that human study outdoor navigation robot. Thorpe et al. [215] present two algorithms, a RGB-based method for road following and a 3D-based method for obstacle detection, for a robot to learn to navigate in a campus. Ross et al. [216] combine deep learning and reinforcement learning to learn obstacle avoidance for UAVs. Morad et al. evaluate the performance of NavACL on two simulated environments, Gibson and Habitat. And We transfer the navigation to a Turtlebot3 wheeled robot (AGV) and a DJI Tello quadrotor (UAV). Both quantitative and qualitative results reveal that the policy of NavACL trained in the simulated environment is surprisingly effective in AGV and UAV. Manderson et al. [217] use conditional imitation learning to train an underwater vehicle to navigate close to sparse geographic waypoints without any prior map.

Many researchers focus on the learning efficiency since the data sampling in the real-world is slow and expensive. Lobos-Tsunekawa et al. [219] propose a map-less visual navigation method for biped humanoid robots. This method extracts information from color images to derive motion commands using using the Deep Deterministic Policy Gradients (DDPG) [77] algorithm. This model runs 20ms on a physical robot, allowing its use in real-time application. Bruce et al. [220] present a method for learning to navigate to a fixed goal on a mobile robot. By using interactive replay of a single traversal of the environment and stochastic environmental augmentation, Bruce et al. demonstrate zero-shot transfer under real-world environmental variations without fine-tuning. To further improve the sample efficient, Pfeiffer et al. [221] leverage prior expert demonstrations for pre-training that reduces training cost. Their model achieves the best performance and shows competitive generalization ability on a real robot platform. Borenstein et al. [222] propose to maintain world model that updated continuously and in real-time to avoid obstacles. The world model learns and simulates the real-world environment and reduce the cost of data sampling [223]. Liu et al. propose Lifelong Federated Reinforcement Learning (LFRL), a learning architecture for navigation in cloud robotic systems to address this problem. This model is able to fuse models and asynchronously evolve the shared model. The architecture has fixed requirements for the dimensions of input sensor signal and the dimensions of action.

Long-range navigation is challenging for real-world robots. Francis et al. [225] present PRM-RL, a hierarchical robot navigation method to tackle this problem. The PRM-RL model consists of a reinforcement learning agents that map noisy sensors to robot controls learn to solve short-range obstacle avoidance tasks, and a sampling-based planner to map the navigation space. Shah et al. [218] propose ViNG, a learning-based navigation system for reaching visually indicated goals and demonstrate this system on a real mobile robot platform. Unlike prior work, ViNG uses purely offline experience and does not require a simulator or online
data collection, which significantly improve the training efficiency. Mapping [226] and path planning [227] is also been widely adopted by many real-world applications. Davison et al. [228] builds an automatic system, which is able to detect, store and track suitable landmark features during goal-directed navigation. They show how a robot uses active vision to provide continuous and accurate global localisation for efficient navigation. Sim et al. [229] enable a robot without active ranger sensors to have the ability of accurately localization by employing a hybrid map representation of 3D point landmarks.

Autonomous driving is another popular field in applying embodied navigation since its remarkable economic values [230]. Most autonomous driving systems from industry today are based on mediated perception approaches [231]. Early works have studied the auto-driving by several computer vision tasks separately, such as car detection and lane detection [232], [233], [234]. Chen et al. [231] propose an autonomous driving framework based on deep neural network to directly perceive RGB inputs and validate its success on the TORCS driving game [235] and real-world data. Caesar et al. [236] propose a new dataset named nuScenes, which comprises 1000 scenes with fully annotated with 3D bounding boxes for 23 classes and 8 attributes. The data scale of nuScenes overwhelms the KITTI dataset to further improve the robustness of detection and tracking. Later, Dosovitskiy et al. [237] introduce CARLA, an open-source simulator for autonomous driving research. Different from the KITTI dataset [238] that provide real-world video for training and testing auto-driving models, CARLA provide a simulator for embodied car navigation. Codevilla et al. [239] train a vehicle by condition imitation learning, which enables the vehicle to responsive to high-level navigational commands. Bansal et al. [240] suggest that behavior cloning is insufficient for handling complex driving scenarios. They propose to synthesized data in the form of perturbations to the expert behavior. Liang et al. [241] present a controllable imitative reinforcement learning (CIRL) framework that explores over a reasonably constrained action space guided by encoded experiences that imitate human demonstrations. Sauer et al. [242] develop intermediate representations for a conditional navigation model to make an agent capable of complex tasks, like navigating intersections, stopping at traffic lights or respecting speed limits.

3.2 Comparison of Simulated and Real Navigation

Today, the simulated navigation problem is still far from the real-world navigation problem. Compared with the simulated environments, the real-world navigation environment is much more complex. An demonstration of the inputs of a widely used simulated environment, Habitat [40], is shown in Fig. 7. An observation of the Habitat environment include a RGB image, a depth image and GPS sensor. The environment contains all static objects. The simulated environments provide unreal synthetic images with fewer objects where the the real-world environments are far more complex with many. Although some simulators [3], [6], [40] provide physical sensors and simulate some physical interactions (such as collision and acceleration), the performance of their physics engine is still far from real. The sensors in the real-world environment, including RGB, GPS and the velocity sensor, are usually noisy while the sensors in the simulated environment have no noise. Lots of obstacles exist during real-world navigation, which blocks the robot from turning or moving forward. Real-world environments are often dynamic since the environment is so complex that many factors are changing in long term or short term, such as temperature, moisture, friction, obstacles and pedestrians. This problem is usually ignored in the simulated navigation tasks. Recently, learning a adaptive policy for a dynamic environment has attracted rising attention. Some work [243], [244], [245] propose simulated robot environments to accomplish this, however, the simulation is far simpler than the real-world.

The domain gap between the simulated and the real settings brings out different challenges, and thus, researchers propose different methods on the two settings respectively. Mobile robot navigation is regarded as a geometric problem, where a robot is required to perceive the geometry of the environment in order to plan collision-free paths towards a goal. The obstacle avoidance is one of the most important challenges and previous works [205], [215], [216], [222], [225] propose many methods to accomplish this. However, robot navigation in simulated tasks are regarded as a policy learning problem that learns a robust navigation policy from a starting position to the target in a complex environment with many possible routes. SLAM-based methods as in [14], [123], [144] contribute a lot to mapping and path planning, which is general for both simulated and real navigation. Deep learning shows its ability in processing images and learning policies for robotic control, which is widely applied in both two settings. However, the usages of deep learning are different. In real-world navigation, the deep neural network is used to perceive RGB inputs [206], predict the future [209] and learn the navigation policy [225]. However, due to the sampling inefficiency and the complex dynamic factors of the real-world environment, the policy is not robust enough. Some works [222], [223] propose to model the environment and other works [225] adopt handcrafted rules to improve the robustness of the navigation policy. In simulated environment, the data sampling is much more efficient. Most
of the simulators render RGB and depth images in more than hundreds of frames per second (FPS), in which the fastest simulator, Habitat [40], achieves 100,000 FPS. Fast data sampling enables the learning with a large batch size. Many works prove that large training batch size leads to robustness in representations. [45], [246]. In spite of the rendered RGB and depth images, some simulated environments are able to provide semantic segmentations [3], [40], [177], a more accurate simulator with few noise to facilitate training. With richer, less noisy data, researchers can apply deeper neural network on navigation agent without worrying about overfitting. For example, transformer [124] is widely applied in navigation works in simulated environments [153], [247] due to its capability of feature representation while it is easily overfitting if it is trained by few noisy data.

### 3.3 Navigation Transferring

Transfer learning is attracting rising attention in embodied navigation. The researchers are motivated from two aspects: 1) learn a navigation agent that is able to perform accurate and efficient navigation in diverse domains and tasks; 2) deploy an agent trained in a simulated environment in a real-world navigation robot. It is challenging to have a model learn skills for navigating in different domains. Moreover, due to the large domain gap between simulated environments and the real-world environment, a well-performed navigation policy trained on a simulated environment cannot be easily transfer to the real-world environment. A lot of navigation tasks have been proposed to research on different capability for navigation in diverse scenarios. In this sections, we discuss the transfer learning in navigation from two different levels: 1) task-level transferring; 2) environment-level transferring, including sim-to-real transferring. The task-level transferring requires the agent to learn a policy that adapts to different input modalities or targets; the environment-level transferring requires the model is invariant to different dynamics and transition functions.

Traoré et al. [249] propose DisCoRL, an approach combining state representation learning and policy distillation for simulated 2D navigation tasks.

Huang et al. [250] transfer pre-trained vision-language representations to solve VLN task by auxiliary losses, Zhang et al. [251] obtain a meta-agent with deep reinforcement learning and then transfers the meta-skill to a robot with a different dimensional configuration using a method named dimension-variable skill transfer. Zhu et al. [63] decompose long navigation instructions into shorter ones, and thus, enables the model to be easily transfer to longer navigation tasks. Chaplot et al. [252] propose a multitask model that jointly learns multimodal tasks, and transferring vision-language knowledge across the tasks. The model adopts a dual-Attention unit to disentangle the vision knowledge and language knowledge and align them with each other. Model-Agnostic mta-learning [253] is widely used in navigation transferring. Li et al. [254] propose an unsupervised reinforcement learning method to learn transferable meta-skills. Wang et al. [255] propose to learn environment-agnostic representations for the navigation policy enables the model to perform on both Vision-Language Navigation (VLN) and Navigation from Dialog History (NDH) tasks. Yan et al. [256] propose MVV-IN, a method acquires transferable transferable meta-skills with multi-modal inputs to cope with new tasks. Liu et al. [257] investigate of how to make robots fuse and transfer their experience so that they can effectively use prior knowledge and quickly adapt to new environments. Gordon et al. [23] propose to decouple the visual perception and policy to facilitates transfer to new environments and tasks.

Sim-real transferring have been well studied in the field of robotic control [258], [259], [260]. Sadeghi et al. [224] firstly propose a learning-based method, as shown in Fig. 9, that train a navigation agent entirely in a simulator and then transfer it into the real world without a single real training image. Consequently, Yuan et al. [261] adopt a sim-real transfer strategy for learning navigation controllers using an end-to-end policy that maps raw pixels as visual input to control actions without any form of engineered feature extraction. Tai et al. [262] trained a robot in simulation with Asynchronous DDPG [77] algorithm and directly deployed the learned controller to the real robot for navigation transferring. Rusu et al. [259] propose to use progressive networks to bridge the reality gap and transfer learned policies for sim-real robot policy transferring. Similarly, Zhu et al. [25] adopt adversarial feature adaptation method that adapt the navigation policy learned from large scale of data in the simulated environment to the real-world environment. Sim-to-real transfer for deep reinforcement learning policies can be applied to complex navigation tasks [263]: including six-legged robots [264], depth-based mapless navigation [265], robots for soccer competitions [266], etc.

### 3.4 Summary

In this section, we introduced the works in real-world navigation application. We compare the difference between simulated environments and real-world environments, and we discover that the difference of environment settings brings the difference of

---

**Fig. 10:** The performances of methods on the Habitat **PointGoal** Challenge, including DD-PPO [89], ego-localization [105], Occupancy Anticipation [248] and SLAM-net.

**Fig. 11:** The performances of methods on the Habitat **ObjectGoal** navigation, including DD-PPO [89], Active Exploration [16], SemExp [17] and 6-Act Tether [267].
challenges to solve. Also, we find that there exist domain gaps between the environments, including the domain gap between different simulated environments and the domain gap between a simulated environment and a real-world environment. Therefore, we introduce some transfer learning works in navigation to give a promising direction to solve this problem.

4 Future Directions

Although lots of works have addressed the navigation problem from diverse aspects, current research progress still far from solving the navigation problem. Current challenges in solving embodied AI could be described into these aspects: 1) The functions and the performance are limited by the embodied environment; 2) the task definitions of navigation has not been fully explored and united; 3) the performances of embodied AI agents in complex environments are still poor; 4) the large domain gap between different tasks and between simulated and real-world environments remains large.

Limitation of the embodied environments. The embodied environments limit the functions of the learned navigation models. If we want a model to equip some functions, we must train the model in an environment with corresponding features. The features could be more diverse input sensors, the specific domain of the data, or larger action space. For instance, compared to the early embodied environments, the large scene in the Matterport3D [2] firstly requires the navigation model to explore and memorize the complex room structure. The vision-language navigation benchmark [8] enables the agents to perceive natural language. The interactive embodied environment like AI2-THOR [38] and iGibson [41] enable the agent to perform interactive actions. The agent learned in an interactive environment is able to move an object, put an object and open a door.

In the future, more functions are needed if we want to obtain a powerful navigation agent. We need the agent to handle a dynamic environment when the objects in the rooms are changing by other agents or humans. In stead of navigating within the navigable areas like [8], [40], we want an agent to find possible roads within a room that has many obstacles. Also, we need a interactive agent that can pick up and put down objects, move chairs and interact with humans. Other modes such as walking, running and climbing are also needed to be considered if we want to build a robust navigator within a complex indoor environment.

The problem definition remains unclear. Even though many embodied navigation tasks and navigation metrics have been proposed, what is a good navigation model is remain unclear. This problem has two folds: 1) what factors have to be considered; 2) how to balance these factors. As we analysed in Sec. 2.2, the accuracy and efficiency are the two main factors to evaluate the performance of navigation. However, in more complex navigation tasks like Help Anna [57], or RMM [190], a questioning-answering setting in the visual dialog navigation, the number of questioning or requesting from the agent are take into consideration. The performance is regarded as lower if the agent ask for more information from human.

Poor navigation performance. Current navigation agents still perform poorly in easy navigation tasks such PointGoal task and ObjectGoal task, as shown in Fig 10 and Fig. 11. The state-of-the-art model performs 64.5% in success rate (SR) and 37.7% in SPL in PointGoal task. And the state-of-the-art model performs 21.08% in success rate (SR) and 8.38% in SPL in ObjectGoal task. As shown in Fig. 4, the current state-of-the-art model [153] in the vision-language navigation task performs 63% in SR and 57% in SPL while the human performance is 86% in SR and 76% in SPL. We conclude that the performances of current models are far from human performance.

Even though previous works have proposed many navigation tasks, they all have their own weaknesses. PointGoal navigation task focuses on perceiving the accumulated forward distance and the relative orientation of an agent while ignores the semantic understanding. ObjectGoal and RoomGoal navigation tasks require an agent to perceive an object in searching while ignores to interact with human via natural language. Also, they are unrealistic if an agent is put in a complex room with many objects and rooms belong to the same class. Vision-language navigation (VLN) [8] asks an agent to reach a target step-by-step by natural language instruction.

Large domain gap. If we want to build a robust navigation agent that is able to work in a real-world environment, we must tackle the problem cause by domain gap. As we analyzed in Sec. 3.2, current simulations cannot fully simulate the real-world environment due to the limit of the rendering technique and physical engine, etc. Therefore, we have to introduce transferring methods to tackle this problem. Even though many transferring methods have been proposed, these methods focus on the transferring of the visual domain or visual feature space. However, there are more challenges to be solve in the real-world environment in spite of the visual domain gap. For instance, the environment is dynamic due to the activity of humans, navigation error due to the error of gears and difference of physical conditions.

5 Conclusion

This paper presented a comprehensive survey on the embodied navigation scenario by summarizing hundreds of works. We thoroughly investigate the environments, tasks and metrics to introduce the problem that the researchers are trying to solve. And we introduce hundreds of methods that solve these tasks in the embodied environments and compare their differences and unique functions. Then we introduce the methods in the real-world environment and demonstrate how the large domain gap led to the drop of navigation performance. At last, we analysis the current problems exist in the embodied navigation and give out four future directions to improve our community.

References

[1] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 190–198.
[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, “Matterport3D: Learning from rgb-d data in indoor environments,” in 2017 International Conference on 3D Vision (3DV), 2017, pp. 667–676.
[3] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, “Gibson env: Real-world perception for embodied agents,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9068–9079.
[4] P. Dourish, Where the action is : the foundations of embodied interaction, 2001.
[5] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, The Description Logic Handbook: Theory, Implementation and Applications, 2003, vol. 32.
[6] M. Savva, A. X. Chang, A. Dosovitskiy, T. A. Funkhouser, and V. Koltun, “Minos: Multimodal indoor simulator for navigation in complex environments.” arXiv preprint arXiv:1712.03931, 2017.
D. S. Chaplot, A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra, “Are we making real progress in simulated environments? measuring the sim2real gap in embodied visual navigation,” in 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3357–3364.

G. Magalhaes, V. Jain, A. Ku, E. Ie, and J. Baldridge, “Effective and general evaluation for instruction conditioned navigation using dynamic time warping,” 2019.

D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, “Objectnav revisited: On evaluation of embodied agents navigating to objects,” arXiv preprint arXiv:2006.13171, 2020.

G. N. DeSouza and A. C. Kak, “Vision for mobile robot navigation: A survey,” IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 2, pp. 237–267, 2002.

T. Kruse, A. K. Pandey, R. Alam, and A. Kirsch, “Human-aware robot navigation: A survey,” Robotics and Autonomous Systems, vol. 61, no. 12, pp. 1726–1743, 2013.

M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan, “Example-based synthesis of 3d object arrangements,” international conference on computer graphics and interactive techniques, vol. 31, no. 6, p. 135, 2012.

A. Meena, M. Savva, O. Maksymets, A. Kadian, and A. R. Zamir, “On evaluation of embodied agents navigating to objects,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2856–2865.

A. A. Rusu, M. A. Zoran, S. Levine, G. S. Sukhbaatar, R. Sukthankar, J. M. Darrell, and A. A. Engel, “Learning to compare and act,” in ICML’14 Proceedings of the 31st International Conference on International Conference on Machine Learning, 2014, pp. 361–369.

D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, “Objectnav revisited: On evaluation of embodied agents navigating to objects,” arXiv preprint arXiv:2006.13171, 2020.

G. N. DeSouza and A. C. Kak, “Vision for mobile robot navigation: A survey,” IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 2, pp. 237–267, 2002.

T. Kruze, A. K. Pandey, R. Alam, and A. Kirsch, “Human-aware robot navigation: A survey,” Robotics and Autonomous Systems, vol. 61, no. 12, pp. 1726–1743, 2013.

M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan, “Example-based synthesis of 3d object arrangements,” international conference on computer graphics and interactive techniques, vol. 31, no. 6, p. 135, 2012.

A. Meena, M. Savva, O. Maksymets, A. Kadian, and A. R. Zamir, “On evaluation of embodied agents navigating to objects,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2856–2865.

A. A. Rusu, M. A. Zoran, S. Levine, G. S. Sukhbaatar, R. Sukthankar, J. M. Darrell, and A. A. Engel, “Learning to compare and act,” in ICML’14 Proceedings of the 31st International Conference on International Conference on Machine Learning, 2014, pp. 361–369.
P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and A. Vakanski, I. Mantegh, A. Irish, and F. Janabi-Sharifi, “Trajectory
Y. Hong, C. Rodriguez-Opazo, Q. Wu, and S. Gould, “Sub-instruction
H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization
D. K. Misra, J. Langford, and Y. Artzi, “Mapping instructions and visual
Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, “Reverie: Remote embodied visual referring expression in real indoor environments,” in ArXiv: Computer Vision and Pattern Recognition, 2019.
L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in European Conference on Computer Vision. Springer, 2016, pp. 69–85.
C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, “Soundspaces: Audio-visual navigation in 3d environments,” in ArXiv: Computer Vision and Pattern Recognition, 2019.
D. K. Misra, J. Langford, and Y. Artzi, “Mapping instructions and visual observations to actions with reinforcement learning,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1004–1015.
H. Chen, A. Suhr, D. Misra, N. Snively, and Y. Artzi, “Touchdown: Natural language navigation and spatial reasoning in visual street environments,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12538–12547.
W. Zhu, H. Hu, J. Chen, H. Deng, V. Jain, E. Le, and F. Sha, “Babywalk: Going farther in vision-and-language navigation by taking baby steps,” ArXiv preprint arXiv:2005.04625, 2020.
D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series,” in KDD workshop, vol. 10, no. 16. Seattle, WA, USA, 1994, pp. 359–370.
H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE transactions on acoustics, speech, and signal processing, vol. 26, no. 1, pp. 43–49, 1978.
A. Vakanski, I. Mantegh, A. Irish, and F. Janabi-Sharifi, “Trajectory learning for robot programming by demonstration using hidden markov model and dynamic time warping,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1039–1052, 2012.
E. J. Keogh and M. Pazzani, “Scaling up dynamic time warping for datamining applications,” in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000, pp. 285–289.
G. Ilharco, V. Jain, A. Ku, E. Le, and J. Baldridge, “General evaluation for instruction conditioned navigation using dynamic time warping,” ViGIL@NeurIPS, 2019.
A. Tamar, Y. Wu, G. Thomas, S. Levine, and P. Abbeel, “Value iteration networks,” in IJCAI’17 Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017, pp. 4094–4099.
L. Lee, E. Parisotto, D. S. Chaplot, E. P. Xing, and R. Salakhutdinov, “Gated path planning networks,” in International Conference on Machine Learning, 2018, pp. 2947–2955.
S. Lenser and M. Veloso, “Visual sonar: fast obstacle avoidance using monocular vision,” in Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No. 03CH37453), vol. 1, 2003, pp. 836–841.
E. Royer, J. Bom, M. Dhome, B. Thuilot, M. Lhuillier, and F. Marmoien, “Outdoor autonomous navigation using monocular vision,” in 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, pp. 1253–1258.
H. Haddad, M. Khatib, S. Lacroix, and R. Chatila, “Reactive navigation in outdoor environments using potential fields,” in Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146), vol. 2, 1998, pp. 1232–1237.
A. Remazeilles, F. Chaumette, and P. Gros, “Robot motion control from a visual memory,” in IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA ’04. 2004, vol. 5, 2004, pp. 4695–4700.
Y. Zhu, D. Gordon, E. Kolve, D. Fox, L. Fei-Fei, A. Gupta, R. Mottaghi, and A. Farhadi, “Visual semantic planning using deep successor representations,” in 2018 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 483–492.
V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Harley, T. P. Lillicrap, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in ICML’16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, 2016, pp. 1925–1935. [Online]. Available: https://papers.nips.cc/paper/2940437961.
T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in ICLR 2016: International Conference on Learning Representations 2016, 2016.
Y. Li and J. Kosecka, “Learning view and target invariant visual servoing for navigation,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 684–685.
A. Mousavian, A. Toshev, M. Fiser, J. Kosecka, A. Wahid, and J. Davidson, “Visual representations for semantic target driven navigation,” in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 8846–8852.
T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European Conference on Computer Vision, 2014, pp. 740–755.
A. Mousavian, H. Pirsiavash, and J. Kosecka, “Joint semantic segmentation and depth estimation with deep convolutional networks,” in 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 611–619.
W. Shen, D. Xu, Y. Zhu, L. Fei-Fei, L. Guibas, and S. Savarese, “Situational fusion of visual representation for visual navigation,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2881–2890.
X. Ye and Y. Yang, “Efficient robotic object search via hiem: Hierarchical policy learning with intrinsic-extrinsic modeling,” ArXiv preprint arXiv:1008.5896, 2020.
Y. Lv, N. Xie, Y. Shi, Z. Wang, and H. T. Shen, “Improving target-driven visual navigation with attention on 3d spatial relationships,” ArXiv preprint arXiv:2005.02153, 2020.
Y. Wu, Y. Wu, A. Tamar, S. Russell, G. Gkioxari, and Y. Tian, “Bayesian relational memory for semantic visual navigation,” in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2769–2779.
Q. Wu, K. Xu, J. Wang, M. Xu, and D. Manocha, “Reinforcement learning based visual navigation with information-theoretic regularization,” ArXiv preprint arXiv:1912.04078, 2019.
A. Stooke and P. Abbeel, “Accelerated methods for deep reinforcement learning,” ArXiv: Learning, 2018.
E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. I. Jordan, and I. Stoica, “Rlib: Abstractions for distributed reinforcement learning,” in 35th International Conference on Machine Learning (ICML), 2018, pp. 3053–3062.
“Almost optimal model-free reinforcement learning via reference-advantage decomposition,” in Advances in Neural Information Processing Systems, vol. 33, 2020.

X. Zhang, Y. Ma, and A. Singla, “Task-agnostic exploration in reinforcement learning,” in Advances in Neural Information Processing Systems, vol. 33, 2020.

T. Chen, S. Gupta, and A. Gupta, “Learning exploration policies for navigation,” in 7th International Conference on Learning Representations, ICLR 2019, 2019.

A. G. Barto and S. Mahadevan, “Recent advances in hierarchical reinforcement learning,” Discrete Event Dynamic Systems, vol. 13, no. 1, pp. 41–77, 2003.

P. Dayan and G. E. Hinton, “Feudal reinforcement learning,” in Advances in Neural Information Processing Systems 5, vol. 5, 1992, pp. 271–278.

E. Kaufmann, M. Gehrig, P. Foehn, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, “Beauty and the beast: Optimal methods meet learning for drone racing,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 690–696.

X. Li, C. Li, Q. Xia, Y. Bisk, A. Celikyilmaz, J. Gao, N. A. Smith, and Y. Choi, “Robust navigation with language pretraining and stochastic adversarial path sampling,” in ECCV (6), 2019, pp. 71–86.

J. Thomason, D. Gordon, and Y. Bisk, “Shifting the baseline: Single modality performance on visual navigation & qa,” arXiv preprint arXiv:1811.00613, 2018.

H. Huang, V. Jain, H. Mehta, J. Baldridge, and E. Le, “Multi-modal discriminative model for vision-and-language navigation,” in Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), 2019, pp. 40–49.

H. Wang, Q. Wu, and C. Shen, “Soft expert reward learning for vision-and-language navigation,” in ECCV (9), 2020, pp. 126–141.

Q. Xia, X. Li, C. Li, Y. Bisk, Z. Sui, J. Gao, Y. Choi, and N. A. Smith, “Multi-view learning for vision-and-language navigation.” arXiv preprint arXiv:2003.00857, 2020.

R. Hu, D. Fried, A. Rohrbach, D. Klein, T. Darrell, and K. Saenko, “Are you looking? grounding to multiple modalities in vision-and-language navigation,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6551–6557.

S. Kurita and K. Cho, “Generative language-grounded policy in vision-and-language navigation with bayes’ rule.” arXiv preprint arXiv:2009.07783, 2020.

J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in Advances in Neural Information Processing Systems, vol. 32, 2019, pp. 13–23.

H. Tan and M. Bansal, “Lxmert: Learning cross-modality encoder representations from transformers,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 5099–5110.

A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, “Improving vision-and-language navigation with image-text pairs from the web,” in ECCV (6), 2020, pp. 259–274.

A. Graves, “Adaptive Wavenet and recurrent neural networks,” arXiv preprint arXiv:1603.09883, 2016.

A. Das, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, “Neural Modular Control for Embodied Question Answering,” in Proceedings of the Conference on Robot Learning (CoRL), 2018.

A. Anand, E. Belilovsky, K. Kastner, H. Larochelle, and A. C. Courville, “Blindfold baselines for embodied qa.” arXiv: Computer Vision and Pattern Recognition, 2018.

Y. Wu, L. Jiang, and Y. Yang, “Revisiting embodiedqa: A simple baseline and beyond.” IEEE Transactions on Image Processing, vol. 29, pp. 3984–3992, 2020.

E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, “Embodied Question Answering in Photorealistic Environments with Point Cloud Perception,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

H. Luo, G. Lin, Z. Liu, F. Liu, Z. Tang, and Y. Yao, “Sceneqa: Video segmentation based visual attention for embodied question answering,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9667–9676.

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017, pp. 2462–2470.

J. Li, S. Tang, F. Wu, and Y. Zhuang, “Walking with mind: Mental imagery enhanced embodied qa,” in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1211–1219.
