Frontier-enhanced Topological Memory with Improved Exploration Awareness for Embodied Visual Navigation

Xinru Cui[1†], Qiming Liu[1†], Zhe Liu[2], and Hesheng Wang[1*]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai, China
[2] MoE Key Lab of Artifcial Intelligence, Al Institute, Shanghai Jiao Tong University, Shanghai, China

Abstract. We present a novel graph memory structure for navigation, called Frontier-enhanced Topological Memory (FTM). Most prior research primarily focuses on maintaining memory representations for explored areas. In contrast, our approach incorporates ghost nodes into the topological map to characterize unexplored but visible regions. The ghost nodes are generated using a geometric method and serve to indicate the geometrically explorable frontiers. We also employ an online-trained implicit representation method to predict perceptual features for ghost nodes based on historical observations. In addition, we develop a Multi-Stage Memory Extraction module (MSME) that can effectively utilize the FTM to extract target-related signals. It focuses particularly on task-specific information and generates actions end-to-end. By using FTM, the agent can improve its capacity for environmental cognition and memory utilization. We evaluate the proposed approach on visual navigation in the photo-realistic Gibson environment. Experimental results conclusively demonstrate that the proposed navigation framework with FTM boosts the agent's exploration awareness and enhances the performance in image-goal navigation tasks. Code is at `https://github.com/IRMVLab/FTM-nav`.

Keywords: Image-goal navigation · Topological memory · Exploration awareness

1 Introduction

In recent years, there has been a notable surge in interest concerning the autonomous navigation of embodied agents [7, 10, 20, 24, 26, 31, 37, 48]. To navigate effectively in large-scale environments, agents must maintain long-term memory representations of the surroundings. Topological maps provide an efficient and concise representation of environmental structures, where nodes symbolize typical landmarks, and edges indicate accessibility and proximity between nodes.

* Corresponding author.
† The first two authors contribute equally.

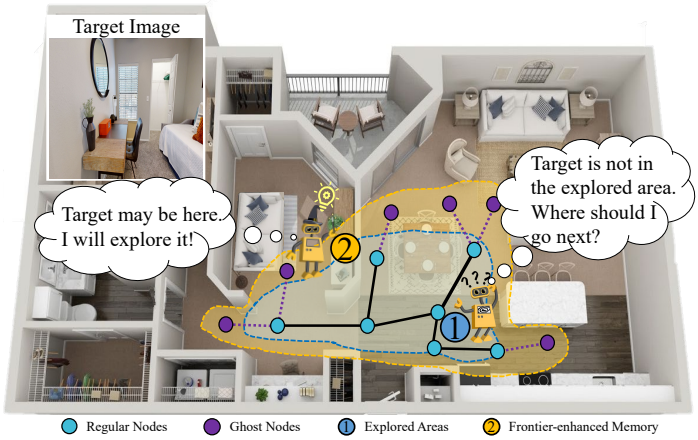


Fig. 1: Design intention of FTM. Most Previous methods solely focus on mapping already explored regions and utilizing these high-confidence scene features. In contrast, our approach integrates ghost nodes to create the Frontier-enhanced Topological Memory, effectively guiding the agent in both exploration and target search.

This inherently offers significant advantages in global planning and strategy generation. Owing to these advantageous properties, numerous learning-based visual navigation frameworks [17, 28, 29, 34] utilize topological maps as a memory mechanism to enhance navigation.

However, existing methods based on topological maps exhibit certain limitations. As shown in Fig. 1, most methods [17, 32] solely map previously explored areas, while failing to represent the broader regions and make full use of boundary features. This limitation leads to inefficient initial exploration in unknown environments. Nevertheless, we believe that extending the topological map to encompass unexplored areas is feasible, as agent's observations often contain visual information about areas that are visible yet to be explored. Therefore, to fully harness geometric structural information from observations and facilitate effective exploration and target search, we introduce *informative ghost nodes* in the topological map to delineate explorable regions.

Some previous studies [11,36] have incorporated virtual nodes into topological maps. However, in these studies, the virtual nodes solely function as directional indicators and lack associated perceptual features, posing challenges in their integration into the memory utilization based on neural networks. Our intention is to equip these ghost nodes with suitable and rich perceptual features, which can not only expand the agent's perceptual scope greatly but also ensure consistent processing of both regular and ghost nodes in subsequent modules. Several studies [14, 15] generate novel views at the boundary of the explored areas by maintaining a 3D point cloud of the environment and employing an image-to-image GAN. Nevertheless, these approaches necessitate extensive pre-training to grasp prior information regarding objects and structures within indoor scenes,

making them unsuitable for integration into our end-to-end framework. Consequently, inspired by the capability of Neural Radiance Field (NeRF) [30] to implicitly and compactly capture environmental features and generate observations of novel perspectives, we incorporate NeRF to extract perceptual features through the neural network implicitly, thereby reconstructing observations at these ghost nodes' positions.

Furthermore, we also focus on designing a feasible structure for action generation. Some studies [4,11,36] tackle this problem by performing path planning in the graph. In contrast, we present a Multi-Stage Memory Extraction structure (MSME) that prioritizes the local spatial distribution of nodes and target cues, ultimately achieving end-to-end action generation.

To summarize, we propose the Frontier-enhanced Topological Memory (FTM) for visual navigation. In our approach, we incorporate ghost nodes to signify geometrically navigable directions. In parallel, we utilize an online-trained NeRF to forecast perceptual features for ghost nodes. We emphasize that NeRF is not used to hallucinate completely unobserved regions. Instead, it reconstructs observations in unvisited but visible boundary areas by capturing scene structure and cues from past observations. The addition of ghost nodes enhances the agent's cognition of the environment and its utilization of boundary features. We also design MSME for end-to-end extraction of task-relevant features and action generation based on FTM.

Our innovations encompass:

- A topological memory structure that integrates informative ghost nodes (rather than merely indicative) to represent unexplored but visible areas and provide guidance for sufficient exploration and target search.
- Endowing the ghost nodes with rich perceptual features based on the online-trained implicit representation, to ensure the consistency of node attributes within the topological graph and subsequent processing.
- A task-specific graph memory extraction mechanism that integrates task state and target cues for adaptive global information fusion and generates actions end-to-end.

2 Related Work

Navigation for Embodied AI. In recent years, learning-based methods have gained significant attention in autonomous driving and target-driven navigation [9,22,25,26,31,46,48]. Substantial advancements have been achieved across various navigation tasks, including image-goal navigation [10, 36], object-goal navigation [2, 3], and vision-and-language navigation [5, 6, 42]. One of the key issues in this field is how to establish a viable memory structure to expand the agent's perceptual scope in time and space, enabling long-term environmental comprehension and reasoning [8]. One approach is to introduce internal memory structures in the network, such as LSTM [31, 37] and Transformer [9, 23]. Another effective strategy involves maintaining an external memory representation of the environment that is independent of the policy network, such as the metric

map [4, 7] and topological map [17, 36]. In this paper, we introduce FTM as an innovative topological memory structure for the environment. Unlike traditional map-based approaches, FTM employs both explicit and implicit memory representations, where implicit memory is utilized to extend the explicit topological map for characterizing unexplored but visible areas.

Navigation with Topological Map. The topological map serves as an abstract and generalized representation of the environment, offering inherent advantages for navigation. Early works only operate on pre-constructed graphs [19, 28, 34], posing challenges when navigating in unknown scenes. Recent studies [13, 17] are dedicated to progressively constructing topological maps online and devising neural-based strategy modules to generate navigation actions. But they primarily concentrate on the construction and update of topological maps in explored regions, often overlooking the significance of features at unexplored but visible areas. Some works [11,36] also introduce virtual nodes in the topological map to signify potential unexplored directions and select the most promising one as the subgoal for navigation. Nonetheless, these ghost nodes solely offer directional information for exploration but lack visual features, limiting these methods to planning-based navigation. In contrast, our approach incorporates ghost nodes to boost agent's exploration awareness and supplies perceptual features for these nodes. It can seamlessly integrate ghost nodes into the designed graph memory extraction structure and enable end-to-end action generation.

Navigation with NeRF. NeRF [30] is impressive for its ability to efficiently produce high-quality representations of complex scene perspectives. There have been several attempts to integrate NeRF into navigation [1, 16, 18, 27, 40]. Some studies aim to establish a transformation relationship between NeRF and explicit spatial geometry representations, generating smooth and collision-free trajectories through planning and optimization [1,16,40], or integrating neural controllers to generate navigation actions [27]. In this paper, the radiance field of the scene is learned dynamically during each episode and serves to endow ghost nodes with explicit perceptual features, achieving topological memory enhancement.

3 Methods

We consider the image-goal navigation task. At each time step, the agent receives first-person RGBD observations. The agent's objective is to reach the image-specified goal location in an unknown environment.

Fig. 2 shows an overview of the proposed framework. The topological map of the environment is constructed incrementally for navigation. We employ a geometric approach to determine the positions of ghost nodes, which represent unvisited regions at the boundaries of explored areas. Additionally, we employ an online-trained NeRF as the implicit memory of the environment to reconstruct visual features for the ghost nodes, enabling their integration into the subsequent memory utilization module MSME.

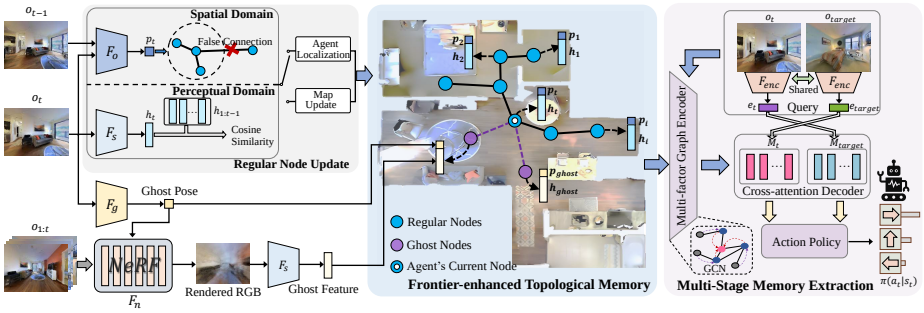


Fig. 2: Pipeline of our proposed method. We introduce Frontier-enhanced Topological Memory (FTM), which dynamically constructs the scene's topological representation during navigation using similarity comparison (F_s) and neural odometry (F_o). FTM incorporates ghost nodes to represent unexplored areas through the geometric method F_g and employs an online-trained NeRF F_n to predict visual features for these ghost nodes. We also design the Multi-Stage Memory Extraction structure (MSME) to extract task-relevant information from FTM and generate actions end-to-end.

3.1 Frontier-enhanced Topological Memory

As shown in Fig. 3, the construction of FTM includes two steps: *Regular Node Update* and *Ghost Node Generation*. The former is responsible for updating the explored areas in the topological map, while the latter focuses on how to generate ghost nodes and predict their visual features.

Regular Node Update. An effective approach to representing the explored areas in the topological map is to evaluate image similarity and correspond the perceptual embedding space to the actual spatial relationship [17, 29, 39]. However, this approximation may not always be reasonable, particularly in cases where a substantial number of repeated features exist in the environment. Thus, we combine similarity estimation and neural odometry to eliminate false edge connections and enhance the reliability of memory construction.

At time t, the nodes within the topological map are denoted as $V = \{v_1, v_2, ..., v_{N_t}\}$. Each node v_i stores observation encoding vector h_i and global pose p_i, where h_i is derived from the pre-trained image encoder F_s based on ResNet-18 [12], and p_i is acquired from the neural odometry F_o [47].

To localize the agent and update the FTM, we first calculate the cosine similarity between the current encoding vector h_t and the vectors in $H = \{h_1, h_2, ..., h_{N_t}\}$, where H stores the node embeddings of the topological map. In the perceptual domain, the current observation is localized in the node v_j with the highest cosine similarity if this value exceeds the threshold s_{th}. Subsequently, we employ the neural odometry F_o for supplementary localization and two-factor validation. The agent's current pose p_t can thus be obtained from F_o. In cases where the distance between v_j and the current position surpasses the threshold d_{th}, we reject the result of similarity localization in v_j. Instead, we use odometry to localize the current observation in the node spatially closest to p_t.

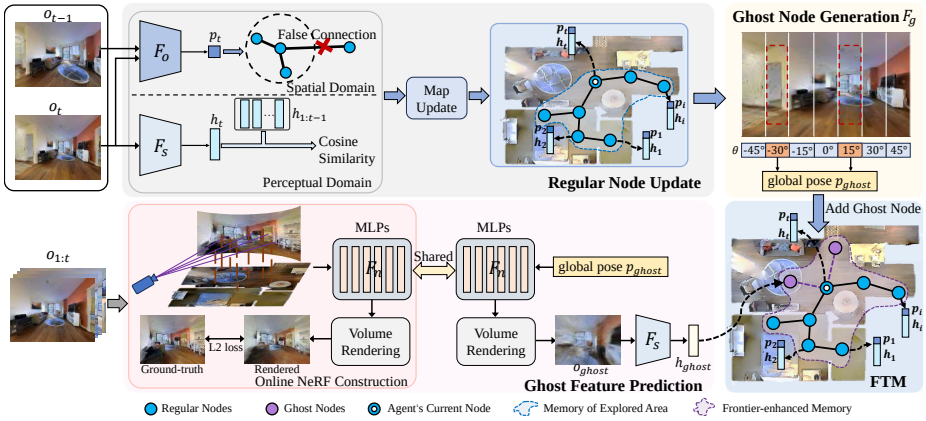


Fig. 3: Construction of FTM. At each timestep, the proposed method updates the explored area in the topological map by combining similarity estimation and visual odometry. It then utilizes depth data to predict geometrically explorable directions and correspondingly adds ghost nodes to the topological map. An online-trained NeRF is employed to characterize scene features and forecast perceptual attributes at virtual nodes to achieve topological memory enhancement.

Upon successful localization, the embedding of the newly localized node is replaced with the current feature h_t, and an edge between the two nodes is established. If the current observation cannot be localized in the established topological map, we create a new node and connect it to the last localized node.

Ghost Node Generation. To enhance the representation of unexplored areas and provide valuable guidance for sufficient exploration and target search, we predict ghost nodes and incorporate them into the topological map to expand the graph. Depth images offer distance information about objects in the field of view, enabling the agent to determine the boundaries of the explorable area and the positions of obstacles. Hence, leveraging agent's current depth observations, we forecast the presence of explorable regions using F_g in seven distinct directions from its current position. F_g evaluates each direction θ within the range of $\{0°, \pm15°, \pm30°, \pm45°\}$, determined by a 90° field of view. We vertically crop the agent's depth observation image to the central 70% area to eliminate interference from the ceiling and floor, and then calculate the average depth in the direction θ. If the average depth exceeds 2.5 meters, we deem the direction explorable. Then the ghost node's relative pose to the current position is represented as (r, θ) where $r = 1m$, and its global pose can be obtained. Subsequently, the ghost node is integrated into the corresponding location within the topological map and connected to the agent's current node. We refrain from introducing a ghost node in the direction θ when a regular node already exists in that direction. Upon the agent's arrival at the ghost node's position, the ghost node is transformed into a regular node and represents the explored areas.

Since we lack visual observations at the positions of ghost nodes, we must reconstruct suitable perceptual information for these nodes based on past observations. Thus, we employ NeRF [30] to compactly represent scene features and render visual observations of ghost nodes triggered by their poses. The implicit representation can be expressed as a mapping function F_n, which maps spatial positions $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ and viewing directions $\mathbf{d} \in \mathbb{R}^2$ to color $\mathbf{c} = (r, g, b)$ and volume density σ.

Then we utilize implicit volume rendering [30] to render the observation o_{ghost} at the ghost node's position. Eqs. (1) and (2) are employed to calculate colors $C(\mathbf{r})$ for every ray $\mathbf{r}$. Each pixel in o_{ghost} corresponds to a ray with near and far bounds t_n and t_f. To optimize computational efficiency, we implement efficient sampling [21] along the ray direction $\mathbf{r}$, thereby facilitating the online training of NeRF.

$$C(r) = \int_{t_n}^{t_f} T(t) \cdot \sigma(\mathbf{r}(t)) \cdot \mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt \tag{1}$$

$$T(t) = \exp\left(- \int_{t_n}^{t} \sigma(\mathbf{r}(s)) \right) ds \tag{2}$$

It is important to note that we do not utilize the intermediate feature vector from F_n as the ghost feature. Instead, we render the observation o_{ghost} at the ghost node's position using F_n. Subsequently, we pass o_{ghost} through the encoder F_s, which is typically used for encoding regular nodes, to generate the ghost feature h_{ghost}. The objective here is to ensure uniform feature distribution among all nodes within the topological map. Additionally, we highlight that the implicit memory representation F_n does not rely on prior knowledge but is dynamically learned from the observation sequence during each episode, which leads to the generality of FTM for scenes.

3.2 Utilizing Memory for Navigation

To extract task-relevant information from FTM, we design the Multi-Stage Memory Extraction (MSME). First, we utilize a Graph Convolutional Network (GCN) to integrate topological information. This process fuses the current observations, target information, and existing perceptual memories, enhancing the agent's understanding of the environmental structure and layout. Then, we employ the cross-attention decoder to extract critical perceptual information from the fused topological map and generate actions end-to-end.

Multi-factor Graph Encoder. Our initial step involves aggregating data from adjacent nodes to potentially broaden the spatial receptive field, concurrently incorporating visual cues for subsequent information extraction. We employ a shared ResNet-18 encoder F_{enc} for encoding both the current observation o_t and the target observation o_{target} into feature vectors e_t and e_{target}. The two vectors are then separately integrated into the topological memory $H = \{h_1, h_2, ..., h_{N_t}\}$. To elaborate further, e_t and e_{target} are respectively concatenated with each node

embedding $h_i \in H$ and then passed through a linear layer FC. This process generates $\mathcal{H}_e^{(0)}$ and $\mathcal{H}_{target}^{(0)}$, fusing information about current and target observations. Then we feed them into a multi-layer GCN to yield multi-factor topological memory M_t and M_{target}. The process can be summarized as follows:

$$\mathcal{H}_{\mathcal{E}}^{(0)} = \mathrm{FC}\left([H, \mathcal{E}]\right), \tag{3}$$

$$\mathcal{H}_{\mathcal{E}}^{(K)} = \mathrm{ReLU}\left(\hat{\mathcal{D}}^{-\frac{1}{2}}\hat{\mathcal{A}}\hat{\mathcal{D}}^{-\frac{1}{2}}\mathcal{H}_{\mathcal{E}}^{(K-1)}W^{(K-1)}\right), \tag{4}$$

$$\mathcal{M} = \mathcal{H}_{\mathcal{E}}^{(L)}, \tag{5}$$

where $\mathcal{E}$ represents e_t or e_{target}, K represents the K-th layer, and $\mathcal{M}$ represents M_t or M_{target} correspondingly. $\hat{\mathcal{A}}$ means the adjacency matrix with self-loops, $\hat{\mathcal{D}}$ corresponds to its degree matrix, and $W \in \mathbb{R}^{d\times d}$ is a parameter matrix.

Cross-attention Decoder. We assert that agents should prioritize task-relevant cues, such as nearby and target environmental information, to enhance spatial awareness and task-related reasoning capabilities. In the structure, attention is reallocated to feature maps M_t and M_{target} to capture relationships between distant nodes.

We employ the multi-head attention function, $mAtt(Q, K, V)$ [41], to extract task-related memories. Specifically, we use e_{target} as the query, and M_t as the key and value to obtain the contextual information x_{target}. Similarly, we acquire x_t by reallocating attention to M_{target} using e_t. The process can be represented as follows:

$$x_{target} = mAtt(e_{target}, M_t, M_t), \tag{6}$$

$$x_t = mAtt(e_t, M_{target}, M_{target}). \tag{7}$$

Action Policy. We then apply residual connections and layer normalization to x_t and x_{target}, and then feeding them into a fully connected layer to obtain y_t and y_{target}:

$$\hat{x}_{target} = \mathrm{LN}\left(x_{target} + e_{target}\right), \tag{8}$$

$$y_{target} = \mathrm{LN}\left(\mathrm{ReLU}\left(\mathrm{FC}\left(\hat{x}_{target}\right)\right) + \hat{x}_{target}\right), \tag{9}$$

where LN represents layer normalization. It's important to highlight that the generation process for both y_t and y_{target} is identical. They share the same attention structures, but their parameters are independently trained.

Finally, y_t, y_{target} and the perceptual features of current observation e_t are concatenated to form the comprehensive cognitive feature, which is finally employed by a MLP and the stacked LSTM to learn the action policy.

4 Experimental Setup

Task Settings. We conduct image-goal navigation in the photorealistic Habitat simulator [35]. The objective is to reach the position of a given target image

within a maximum timestep $N_m = 300$ while avoiding obstacles. The robot carries an RGBD camera in the body front with 144×192 resolution and $90°$ HFOV. In line with previous methods, we incorporate real actuation motion noise and utilize a discrete action space with four types of actions: move forward ~ 0.25m, turn right $\sim 15°$, turn left $\sim 15°$, and stop. The navigation is considered successful when the robot reaches the target within $1m$.

Training Process and Test Settings. We first train the image encoder F_s and the neural odometer F_o using the approaches in [17] and [47]. Then we freeze the parameters of F_s and F_o, and employ multithreading to train the navigation policy. In the primary thread, the agent incrementally constructs the FTM, which the MSME module subsequently uses to generate actions. These actions are optimized through imitation learning (IL) to minimize the disparity with the optimal actions. Another thread focuses on the online training of F_n during navigation. F_n is dynamically learned during each episode based on the photometric loss and the depth loss between the rendered images and the real observations. The training dataset comprises 50 highly realistic indoor environments on the Gibson dataset [44]. Each episode in the training set represents the optimal path from the starting point to the target, obtained through expert control.

During testing, we employed a separate set of 9 environments, distinct from those used for training. The test dataset consists of 810 episodes, categorized into three levels of difficulty based on the distance between the starting point and the target: 1) Easy: $1.5m - 3.0m$; 2) Medium: $3.0m - 5.0m$; 3) Hard: $5.0m - 10.0m$.

Evaluation Metrics. We use three metrics to evaluate the navigation performance: *Success Rate (SR)*. This is the ratio between the the successful and the total number of episodes. *Success weighted by Path Length (SPL)*. This estimates the path efficiency along with the success rate. *Distance To Success (DTS)*. This measures the closest distance of the agent from the success threshold boundary in the episodes.

Baseline Models. We compare the proposed method with the following baselines: 1) **Reactive** [48]: The robot navigates in the environment without any memory, i.e., the robot extracts perception features and then directly generates actions using MLP; 2) **Nav A3C** [31]: An internal-memory-based method that maintains memory using LSTM; 3) **OVRL+SLING** [43, 45]: A reinforcement learning-based method with pretrained visual representation. It also combines a last-mile navigation method for efficient target reaching; 4) **MSM** [33]: It combines short-term working memory and long-term episodic memory to generate actions; 5) **NRNS** [11]: This method uses a topological map and select nodes to explore through a distance estimator. Distance function and target prediction function are learnt over passive videos of roaming; 6) **VGM** [17]: It evaluates the similarity of observed images to construct a topological visual memory, and then generate actions using GNNs; 7) **FGPrompt** [38]: It proposes a novel fine-grained goal prompting architecture to learn observation embeddings conditioned on the fine-grained and high-resolution features of the goal image.

Ablation Models. We ablate three crucial modules to verify their roles in navigation: *ghost node prediction* (F_g), *ghost feature prediction* (F_n), and *MSME*.

Table 1: Comparison of navigation performance with baselines.

	Easy			Medium			Hard			Overall		
	SR↑	**SPL↑**	**DTS↓**	**SR↑**	**SPL↑**	**DTS↓**	**SR↑**	**SPL↑**	**DTS↓**	**SR↑**	**SPL↑**	**DTS↓**
Reactive [48]	0.500	0.332	0.527	0.217	0.129	1.677	0.106	0.062	4.087	0.274	0.174	2.097
Nav A3C [31]	0.608	0.301	0.325	0.410	0.186	0.923	0.124	0.047	3.023	0.381	0.178	1.424
MSM [33]	0.616	0.342	0.403	0.511	0.323	0.767	0.294	0.146	2.405	0.474	0.270	1.192
NRNS [11]	0.521	0.364	0.718	0.369	0.276	1.385	0.200	0.125	2.570	0.363	0.255	1.558
VGM [17]	0.644	0.385	0.288	0.575	**0.373**	0.671	0.384	0.225	1.880	0.534	0.328	0.946
OVRL+SLING [43, 45]	0.702	0.479	0.317	0.590	0.325	0.760	0.416	0.220	2.006	0.569	0.341	1.027
FGPrompt [38]	**0.796**	0.463	0.293	0.562	0.291	0.682	0.403	0.206	1.879	0.587	0.320	0.951
FTM(Ours)	0.774	**0.485**	**0.185**	**0.650**	0.360	**0.548**	**0.451**	**0.252**	**1.523**	**0.625**	**0.366**	**0.752**

Ablating F_g implies that we only maintain memory representations for explored areas, without introducing ghost nodes to the topological map. When F_n is ablated, the features of ghost nodes are not predicted using NeRF. Instead, these features are set to zero vectors, indicating that ghost nodes solely offer directional information for exploration but lack visual attributes. Ablating MSME involves only using simple multi-layer GCNs and self-attention to extract the raw topological memory H for navigation, without employing the multi-factor graph encoder and the cross-attention decoder.

5 Results

We evaluate the navigation performance of various models in different test scenes. Furthermore, we conduct ablation and interpretability experiments to demonstrate the effectiveness of the proposed navigation framework.

5.1 Navigation Performance.

Comparison with Baselines. Tab. 1 shows the performance of different models in various difficulty settings. In contrast to the *Reactive* model, which lacks memory mechanisms, and the models employing implicit memory mechanisms (*Nav-A3C* and *OVRL+SLING*), our model demonstrates superior navigation performance, especially in the hard setting. We assert that employing a concise explicit representation of the environment and finding how to enhance exploration awareness based on such representation are crucial.

Compared to other baseline models with topology-based memory structures (*VGM* and *NRNS*), our approach achieves significant performance gains on the overall metrics. In contrast to *VGM*, our inclusion of ghost nodes within the topological map for representing unexplored areas significantly broadens the agent's perceptual field, thus facilitating exploration and target search. Compared to *NRNS*, we endow the ghost nodes with perceptual features through an online-trained NeRF. This ensures that the ghost nodes are integrated into the memory utilization module to make full use of the boundary features and achieve topological enhancement.

Table 2: The ablation studies on different components.

			Easy		Medium		Hard		Overall	
F_g	F_n	MSME	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**
✗	✗	✓	0.692	0.406	0.593	0.356	0.325	0.193	0.537	0.318
✓	✗	✓	0.702	0.414	**0.652**	0.359	0.376	0.195	0.577	0.323
✓	✓	✗	0.731	0.437	0.572	0.333	0.345	0.207	0.549	0.325
✓	✓	✓	**0.774**	**0.485**	0.650	**0.360**	**0.451**	**0.252**	**0.625**	**0.366**

Table 3: Comparisons on different ghost feature types. "Previous view" indicates the ghost features are extracted from the partial view at the previous node. "SE3DS" applies an image-to-image GAN to generate the visual observation for the ghost node. "G.T. observation" represents the ground truth observation of the ghost node.

	Easy		Medium		Hard		Total	
Ghost Feature Type	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**	**SR↑**	**SPL↑**
Previous view	0.732	0.392	0.635	0.337	0.382	0.201	0.583	0.310
SE3DS [14]	0.752	0.397	<u>0.664</u>	<u>0.361</u>	0.423	0.225	0.613	0.328
Ours (NeRF)	<u>0.774</u>	**0.485**	0.650	0.360	<u>0.451</u>	<u>0.252</u>	<u>0.625</u>	<u>0.366</u>
G.T. observation	**0.782**	<u>0.482</u>	**0.669**	**0.371**	**0.496**	**0.290**	**0.649**	**0.381**

Furthermore, our FTM module is designed to capture boundary and low-confidence scene features, enhancing the agent's active exploration and information gathering. Consequently, our method may achieve suboptimal performance in simpler settings where the agent can easily see the target and requires minimal exploration, but it excels in more challenging settings and overall performance.

Ablation Study. Tab. 2 presents the detailed ablations of the proposed method. (1) The absence of the MSME module significantly reduces navigation performance, demonstrating its effectiveness in integrating multi-factor information from targets and current cues to guide the agent in extracting task-specific signals. (2) Compared to the model without ghost nodes (F_g), our *full* model increases SR by 16.4% and SPL by 15.1%. we speculate that the addition of ghost nodes enhances the initial exploration, allowing the agent to detect target clues earlier and consequently improving performance. (3) In comparison to the model without ghost features (F_n), our *Full* model performs better under hard settings, showing a 19.9% improvement on success rate and a 29.2% improvement on SPL. This is unsurprising because as navigation difficulty increases, the agent with reasonable predictions about the unexplored environment is more likely to choose the promising directions for exploration and avoid redundant searches in specific regions. We believe that the ghost node features predicted through F_n and F_s indeed bolster the agent's comprehension of uncharted areas, enabling the full utilization of the boundary features.

To further explore the effectiveness of the visual features extracted from NeRF reconstructions for ghost nodes, we consider different types of ghost features in Tab. 3. Compared to ghost features extracted from the partial view at the previous node or from the generative model, our proposed NeRF-based ap-

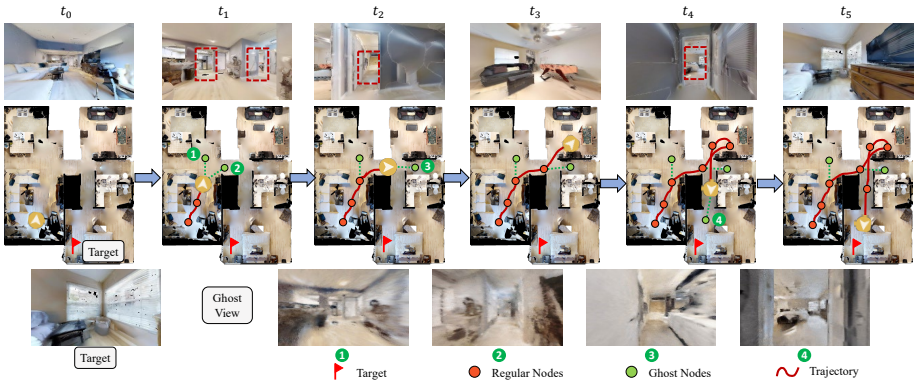


Fig. 4: Example visualization of a typical navigation episode. To visualize the agent's behavior, we plot its trajectory on the top-down map of the environment. Additionally, we present the agent's actual observations at each time step, as well as the rendered observations at the ghost nodes. Our method effectively prompts the agent to fully explore and search for targets in unknown scenes.

proach demonstrates a significant performance increase. This suggests that our method excels at capturing boundary features in unexplored but visible areas based on historical observations, potentially enhancing overall navigation task efficiency. Furthermore, we attach the ground truth observations obtained from the simulator to the ghost nodes. The results indicate that utilizing visual encodings derived from NeRF reconstructions as ghost features results in only a minor decline in navigation performance (3.7% in SR and 3.9% in SPL) when compared to the ground truth version. This further confirms the effectiveness of integrating NeRF to predict these ghost features.

5.2 Visualization and Interpretability Experiments

Behavior Analysis in Typical Scene. Fig. 4 shows the trajectory of the agent and the observations rendered at ghost nodes. It is evident that the agent constructs its environmental understanding during navigation, and is guided by the ghost nodes in exploration. In the initial phase of the task, the agent exhibits limited environmental awareness and tends to explore unfamiliar areas while avoiding collisions under the guidance of ghost nodes. At t_3, the agent recognizes that the target is not within the current room and exits. At t_4, it explores in the direction of the bedroom indicated by the ghost node, efficiently avoiding redundant paths and seeking the target point along new feasible routes. Upon receiving a similar image to the target at t_5, the agent decisively advances towards that area.

Node Attention Visualization. Fig. 5 illustrates the FTM construction process and the allocation of node attention during different navigation episodes. The cross-attention decoder prioritizes two types of nodes: *ghost nodes* which

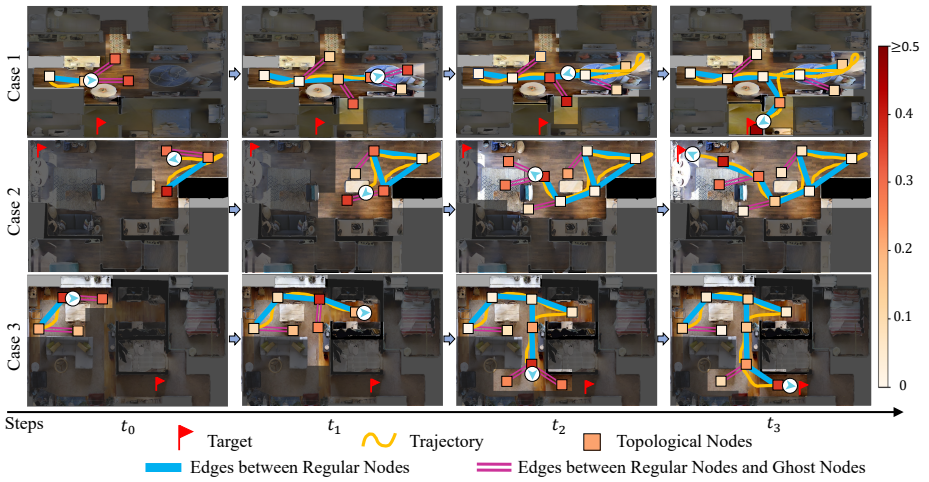


Fig. 5: Visualization of the node attention in various episodes. The rectangle represents a node in the topological map, with its color indicating the attention score for x_{target}. This score reflects the agent's focus on the target information. A darker color signifies a higher level of attention on that node. Ghost nodes and nodes close to the agent typically receive higher attention scores during navigation.

are more likely to lead to the target point, and *nodes near the agent's current location*. In the initial stages of navigation, we notice that the agent primarily focuses its attention on the ghost nodes. This suggests that the agent prioritizes unexplored areas early in the process, formulating actions to fully explore the scene guided by these ghost nodes.

Interestingly, the inclusion of ghost nodes and the MSME module may prevent the agent from easily encountering a deadlock in a particular room. For example, in time step t_1 of case 3, the agent discovers that the target is not in the current room and decides to leave. At this point, previously visited nodes or ghost nodes are assigned higher attention scores because their features may resemble the target observations more closely. Subsequently, the agent returns to these nodes, and resumes the target search under the guidance of ghost nodes until it reaches the destination. We argue that both ghost nodes and the MSME module effectively encourage agent's sufficient exploration and direct it to prioritize task-specific information in unkown environments.

5.3 Runtime Analysis

We use the NeRF acceleration toolbox [21] and parallelize our system into two threads to speed up NeRF optimization: one for NeRF training and the other for navigation. For images sized 144×192, each NeRF training iteration takes 5.6ms. The agent's action time step is 0.1s, allowing NeRF to optimize for 18 iterations per frame. This enables online and real-time NeRF training during navigation.



Fig. 6: Real robot experimental results. We present the global and camera views during the image-goal navigation process, along with the robot's trajectory.

At the start of each episode, we additionally optimize NeRF for the first three frames with 300 iterations (1.7s) to achieve a good initialization. Experiments are conducted on two NVIDIA RTX 4090 GPUs.

5.4 Real-world Experiments

We deploy our method on a wheeled differential robot in the real world to validate its adaptability and feasibility for real-time usage. The experimental site is an outdoor corridor with scattered obstacles. The robot is equipped with an Intel D435i camera and an NVIDIA Jetson AGX Xavier platform. As illustrated in Fig. 6, the robot successfully navigates by stopping at the correct goal location, demonstrating effective obstacle avoidance and target understanding. Furthermore, although our method is trained in static scenes, we find that its obstacle avoidance capability generalizes to dynamic obstacles and pedestrians in real-world testing (*e.g.*, the robot will bypass pedestrians and continue its goal search). We believe the robot has learned to use depth data to perceive distances and avoid collisions.

6 Conclusion and Limitation

In this paper, we have proposed FTM to boost the agent's exploration awareness and enhance its long-term task understanding. The FTM is constructed online, and ghost nodes are incorporated to signify unexplored regions. An online-trained NeRF is employed to endow these ghost nodes with perceptual features, augmenting the robot's environmental comprehension and maintaining consistency in node feature distribution. We show that FTM helps the agent focus on geometrically explorable areas and evaluate the likelihood of these areas leading to the target like humans, which is advantageous for target-driven navigation. However, the proposed method still has limitations. The implicit memory serves solely to generate visual features for ghost nodes, lacking effective integration with the explicit memory representation. An interesting direction involves leveraging implicit memory to optimize pose estimation and aid the agent in localization within topological maps.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China under Grant 62225309, 62073222, U21A20480 and 62361166632.

References

1. Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., Schwager, M.: Vision-only robot navigation in a neural radiance world. IEEE Robotics and Automation Letters **7**(2), 4606–4613 (2022). https://doi.org/10.1109/LRA.2022.3150497
2. Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., Wijmans, E.: Objectnav revisited: On evaluation of embodied agents navigating to objects. ArXiv **abs/2006.13171** (2020), https://api.semanticscholar.org/CorpusID:219981280
3. Chaplot, D.S., Gandhi, D., Gupta, A., Salakhutdinov, R.: Object goal navigation using goal-oriented semantic exploration. In: In Neural Information Processing Systems (2020)
4. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. In: International Conference on Learning Representations (ICLR) (2020)
5. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11271–11281 (2021). https://doi.org/10.1109/CVPR46437.2021.01112
6. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16516–16526 (2022). https://doi.org/10.1109/CVPR52688.2022.01604
7. Chen, T., Gupta, S., Gupta, A.: Learning exploration policies for navigation. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=SyMWn05F7
8. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied ai: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence **6**(2), 230–244 (2022). https://doi.org/10.1109/TETCI.2022.3141105
9. Fang, K., Toshev, A., Fei-Fei, L., Savarese, S.: Scene memory transformer for embodied agents in long-horizon tasks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 538–547 (2019). https://doi.org/10.1109/CVPR.2019.00063
10. Gupta, S., Davidson, J., Levine, S., Sukthankar, R., Malik, J.: Cognitive mapping and planning for visual navigation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7272–7281 (2017). https://doi.org/10.1109/CVPR.2017.769
11. Hahn, M., Chaplot, D.S., Tulsiani, S., Mukadam, M., Rehg, J.M., Gupta, A.: No rl, no simulation: Learning to navigate without navigating. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 26661–26673. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/e02a35b1563d0db53486ec068ebab80f-Paper.pdf

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
13. Kim, N., Kwon, O., Yoo, H., Choi, Y., Park, J., Oh, S.: Topological semantic graph memory for image-goal navigation. In: 6th Annual Conference on Robot Learning (2022), https://openreview.net/forum?id=xjTUxBfIzE
14. Koh, J.Y., Agrawal, H., Batra, D., Tucker, R., Waters, A., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: Simple and effective synthesis of indoor 3d scenes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1169–1178 (2023)
15. Koh, J.Y., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: Pathdreamer: A world model for indoor navigation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14718–14728 (2021). https://doi.org/10.1109/ICCV48922.2021.01447
16. Kurenkov, M., Potapov, A., Savinykh, A., Yudin, E., Kruzhkov, E., Karpyshev, P., Tsetserukou, D.: Nfomp: Neural field for optimal motion planner of differential drive robots with nonholonomic constraints. IEEE Robotics and Automation Letters **7**(4), 10991–10998 (2022). https://doi.org/10.1109/LRA.2022.3196886
17. Kwon, O., Kim, N., Choi, Y., Yoo, H., Park, J., Oh, S.: Visual graph memory with unsupervised representation for visual navigation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15870–15879 (2021). https://doi.org/10.1109/ICCV48922.2021.01559
18. Lee, S., Chen, L., Wang, J., Liniger, A., Kumar, S., Yu, F.: Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. IEEE Robotics and Automation Letters **7**(4), 12070–12077 (2022). https://doi.org/10.1109/LRA.2022.3212668
19. Li, D., Zhang, Q., Zhao, D.: Graph attention memory for visual navigation. In: 2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS). pp. 1–7 (2022). https://doi.org/10.1109/DOCS55193.2022.9967733
20. Li, F., Guo, C., Luo, B., Zhang, H.: Multi goals and multi scenes visual mapless navigation in indoor using meta-learning and scene priors. Neurocomputing **449**, 368–377 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2021.03.084, https://www.sciencedirect.com/science/article/pii/S0925231221004707
21. Li, R., Gao, H., Tancik, M., Kanazawa, A.: Nerfacc: Efficient sampling accelerates nerfs. arXiv preprint arXiv:2305.04966 (2023)
22. Liu, J., Wang, G., Jiang, C., Liu, Z., Wang, H.: Translo: A window-based masked point transformer framework for large-scale lidar odometry. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1683–1691 (2023)
23. Liu, J., Wang, G., Liu, Z., Jiang, C., Pollefeys, M., Wang, H.: Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8451–8460 (2023)
24. Liu, J., Wang, G., Ye, W., Jiang, C., Han, J., Liu, Z., Zhang, G., Du, D., Wang, H.: Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15109–15119 (2024)
25. Liu, J., Zhuo, D., Feng, Z., Zhu, S., Peng, C., Liu, Z., Wang, H.: Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment. arXiv preprint arXiv:2403.18274 (2024)

26. Liu, Z., Liu, Q., Tang, L., Jin, K., Wang, H., Liu, M., Wang, H.: Visuomotor reinforcement learning for multirobot cooperative navigation. IEEE Transactions on Automation Science and Engineering **19**(4), 3234–3245 (2022). https://doi.org/10.1109/TASE.2021.3114327
27. Marza, P., Matignon, L., Simonin, O., Wolf, C.: Multi-object navigation with dynamically learned neural implicit representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11004–11015 (October 2023)
28. Meng, X., Ratliff, N., Xiang, Y., Fox, D.: Scaling local control to large-scale topological navigation. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 672–678 (2020). https://doi.org/10.1109/ICRA40945.2020.9196644
29. Mezghan, L., Sukhbaatar, S., Lavril, T., Maksymets, O., Batra, D., Bojanowski, P., Alahari, K.: Memory-augmented reinforcement learning for image-goal navigation. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3316–3323 (2022). https://doi.org/10.1109/IROS47612.2022.9981090
30. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
31. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., Hadsell, R.: Learning to navigate in complex environments. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=SJMGPrcle
32. Niwa, T., Taguchi, S., Hirose, N.: Spatio-temporal graph localization networks for image-based navigation. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3279–3286 (2022). https://doi.org/10.1109/IROS47612.2022.9981958
33. Sang, H., Jiang, R., Wang, Z., Zhou, Y., He, B.: A novel neural multi-store memory network for autonomous visual navigation in unknown environment. IEEE Robotics and Automation Letters **7**(2), 2039–2046 (2022). https://doi.org/10.1109/LRA.2022.3140795
34. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=SygwwGbRW
35. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A platform for embodied ai research. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9338–9346 (2019). https://doi.org/10.1109/ICCV.2019.00943
36. Singh Chaplot, D., Salakhutdinov, R., Gupta, A., Gupta, S.: Neural topological slam for visual navigation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12872–12881 (2020). https://doi.org/10.1109/CVPR42600.2020.01289
37. Singla, A., Padakandla, S., Bhatnagar, S.: Memory-based deep reinforcement learning for obstacle avoidance in uav with limited environment knowledge. IEEE Transactions on Intelligent Transportation Systems **22**(1), 107–118 (2021). https://doi.org/10.1109/TITS.2019.2954952
38. Sun, X., Chen, P., Fan, J., Chen, J., Li, T.H., Tan, M.: FGPrompt: Fine-grained goal prompting for image-goal navigation. In: Thirty-seventh Conference on Neu-

ral Information Processing Systems (2023), https://openreview.net/forum?id=oaJEB5Qcia
39. Taniguchi, A., Sasaki, F., Yamashina, R.: Pose invariant topological memory for visual navigation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15364–15373 (2021). https://doi.org/10.1109/ICCV48922.2021.01510
40. Tong, M., Dawson, C., Fan, C.: Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10511–10517 (2023). https://doi.org/10.1109/ICRA48891.2023.10161482
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
42. Wang, H., Wang, W., Liang, W., Xiong, C., Shen, J.: Structured scene memory for vision-language navigation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8451–8460 (2021). https://doi.org/10.1109/CVPR46437.2021.00835
43. Wasserman, J., Yadav, K., Chowdhary, G., Gupta, A., Jain, U.: Last-mile embodied visual navigation. In: 6th Annual Conference on Robot Learning (2022), https://openreview.net/forum?id=RgJwDQwW82y
44. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9068–9079 (2018). https://doi.org/10.1109/CVPR.2018.00945
45. Yadav, K., Ramrakhya, R., Majumdar, A., Berges, V.P., Kuhar, S., Batra, D., Baevski, A., Maksymets, O.: Offline visual representation learning for embodied navigation. In: Workshop on Reincarnating Reinforcement Learning at ICLR 2023 (2023), https://openreview.net/forum?id=Spfbts_vNY
46. Zhang, M., McCarthy, Z., Finn, C., Levine, S., Abbeel, P.: Learning deep neural network policies with continuous memory states. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 520–527 (2016). https://doi.org/10.1109/ICRA.2016.7487174
47. Zhao, X., Agrawal, H., Batra, D., Schwing, A.: The Surprising Effectiveness of Visual Odometry Techniques for Embodied PointGoal Navigation. In: Proc. ICCV (2021)
48. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 3357–3364 (2017). https://doi.org/10.1109/ICRA.2017.7989381