2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

HSPNav: Hierarchical Scene Prior Learning for Visual Semantic Navigation Towards Real Settings

Jiaxu Kang*, Bolei Chen*, Ping Zhong†, Haonan Yang, Yu Sheng, and Jianxin Wang

***Abstract*— Visual Semantic Navigation (VSN) aims at navigating a robot to a given target object in a previously unseen scene. To tackle this task, the robot must learn a nimble navigation policy by utilizing spatial patterns and semantic co-occurrence relations among objects in the scene. Prevailing approaches extract scene priors from the instant visual observations and solidify them in neural episodic memory to achieve flexible navigation. However, due to the oblivion and underuse of the scene priors, these methods are plagued by repeated exploration, effective-knowledge sparsity, and wrong decisions. To alleviate these issues, we propose a novel VSN policy, HSPNav, based on Hierarchical Scene Priors (HSP) and Deep Reinforcement Learning (DRL). The HSP contains two components, i.e., the egocentric semantic map-based Local Scene Priors (LSP) and the commonsense relational graph-based Global Scene Priors (GSP). Then, efficient semantic navigation is achieved by employing an immediate LSP to retrieve conducive contextual memories from the GSP. By utilizing the MP3D dataset, the experimental results in the Habitat simulator demonstrate that our HSP brings a significant boost over the baselines. Furthermore, we take an essential step from simulation to reality by bridging the gap from Habitat to ROS. The migration evaluations show that HSPNav can generalize to realistic settings well and achieve promising performance.**

I. INTRODUCTION

Visual **S**emantic **N**avigation (VSN) task [1], [2] aims at navigating a robot to a given object category based on its visual observation in a previously unseen scene. Although unknown scenes imply unfamiliarity and challenge, it is exceedingly rare that no prior information is available in domestic scenes. For example, intelligent bodies like humans often rely on experience to infer the subordination and co-occurrence relations among objects in domestic scenes, which is leveraged to assist in VSN. In other words, humans benefit from traversing a wide variety of domestic scenes over time, summarizing and memorizing a wealth of scene priors that aid in localizing and navigating to specified objects.

In recent years, large-scale realistic 3D scene datasets [3], [4] have provided the impetus for the development of VSN. Existing VSN methods [5]–[14] have attempted to exploit the scene priors by developing implicit **N**eural **E**pisodic **M**emory (NEM) [5], [6] and explicit semantic map representation [7]–[14]. Despite the promising achievements

This work was supported in part by the National Natural Science Foundation of China under 62172443 and in part by the Natural Science Foundation of Hunan Province under 2022JJ30760.

The authors are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China.
† Corresponding Author (e-mail: ping.zhong@csu.edu.cn).
* These authors contributed equally to this work.

The code and video is published at: https://sites.google.com/view/hspnav

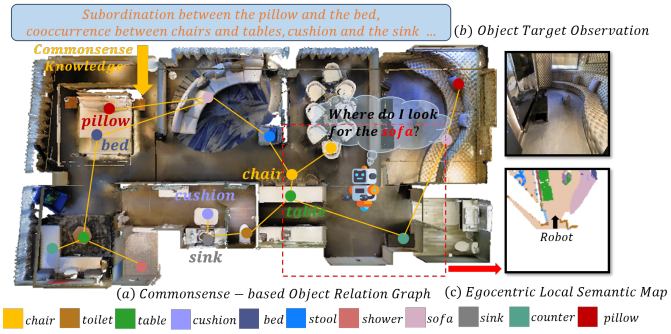

Fig. 1. illustrates the HSP consisting of an egocentric local semantic map-based LSP and a commonsense relational graph-based GSP. The LSP provides robots with fine-grained spatial and semantic features around them. The GSP provides robots with commonsense knowledge and high-level environmental topological information that benefit VSN.

achieved, due to the oblivion and underuse of the scene priors, many methods are plagued by repeated exploration, effective-knowledge sparsity, and wrong decisions. On the one hand, VSN is usually a long-sequence decision task, while recurrent neural networks are known to be inefficient in capturing long-term dependencies [15], [16], which results in repeatedly exploring and failing to execute serialized action. On the other hand, cutting-edge research [1] has verified that semantic map-based environmental representation facilitates the migration from simulation to reality and is more robust. However, the scene priors provided by semantic maps are still raw and sparse, which requires further feature mining. Despite a recent work [17] suggesting the usage of visual language models [18] to enhance the extraction of semantic relations, efficient scene-prior learning remains an open and challenging topic.

In this work, a **H**ierarchical **S**cene **P**riors (HSP) learning method is proposed for VSN inspired by the human habit of summarizing and exploiting experiences, as shown in Fig. 1. In particular, the HSP contains two components, i.e., the egocentric semantic map-based **L**ocal **S**cene **P**riors (LSP) and the commonsense relational graph-based **G**lobal **S**cene **P**riors (GSP). An egocentric local semantic map is employed to support the robot's prompt response to surrounding emergencies, such as collision avoidance and local semantic inference. The GSP is achieved by injecting human commonsense knowledge and objects' texture features into an object-oriented topological relational graph. We humans are adept at responding to current challenges by recalling past experiences in light of current encounters. Inspired by this, a novel HSP-based VSN strategy named HSPNav is further proposed in the **D**eep **R**einforcement **L**earning (DRL) framework by employing an immediate LSP to retrieve

conducive contextual memories from the GSP.

Sufficient comparative studies in the Habitat simulator [19] demonstrate that our method outperforms several baselines. Furthermore, we take an essential step from simulation to reality by bridging the gap from the Habitat simulator to **R**obot **O**perating **S**ystem (ROS) to generalize our approach to real scenes. Qualitative migration evaluations demonstrate the potential of our method to be applied in realistic scenes. Overall, our main contributions are summarized as follows: **(1)** An efficient hierarchical scene priors learning technique and a novel DRL-based VSN strategy HSPNav are proposed. **(2)** An interface is designed to bridge the Habitat simulator and ROS to generalize our HSPNav to more realistic situations. **(3)** Sufficient comparative studies and qualitative evaluations validate the superiority of our method. We will open-source our experimental code and the Habitat-ROS simulation environments for the benefit of the community.

II. Related Work

A. Scene Prior Learning for Visual Semantic Navigation

Mainstream VSN strategies typically learn scene priors by extracting implicit NEM directly from visual observations or by building explicit environmental representations. RedRabbit [5] predicts the next-best action from the RGBD images and obtains more abundant scene priors by introducing auxiliary tasks. SAVN [6] introduces meta-learning skills into the model to empower agents to adapt to unknown environments based on implicit NEM. Although it is succinct to learn navigational skills directly from raw visual images in an end-to-end manner, such approaches are prone to lead to the oblivion and underuse of the scene priors. To remedy these deficiencies, episodic semantic map-based explicit scene representation has been proposed with promising achievements [7]–[14]. SemExp [7] utilizes differentiable projection operations to map the RGBD visual observations into a global semantic map for GSP learning. Then, SemExp plans a collision-free path toward a long-term navigation goal based on the learned GSP. Subsequently, SSCNav [8] exploits the egocentric local semantic map completed by a confidence-aware completion network to infer the next best action. SSCNav demonstrates that fine-grained local semantic maps provide valuable scene priors.

Recently, PONI [9] proposes to predict potential functions to balance the exploration and exploitation of semantic map-based scene priors for efficient VSN. Despite the promising progress, the cumulative construction of semantic maps inevitably results in a significant increase in computation. Therefore, how to mine and exploit high-quality features that contribute to VSN remains an open topic. Most recently, ZSON [17] exploits the pre-trained vision and language model CLIP [18] to empower the agent to imagine the appearance and locate the position of the target object. In this work, we propose to learn hierarchical scene priors by developing a fine-grained local semantic map-based LSP and a commonsense relational graph-based GSP, respectively. Inspired by human navigational skills, we design an innovative scene prior retrieval mechanism to fully leverage the HSP.

B. Visual Semantic Navigation Towards Real Settings

Although semantic map-based VSN strategies have been validated to generalize to real scenes, the migration between simulation and reality still faces many challenges. A cutting-edge study [1] in the VSN field suggests that improving the visual quality of 3D simulation scans is an essential means to reduce the domain gap. Some methods exploit adversarial training [20] and domain randomization [21] to perform the domain transfer between the real and simulated environments in terms of image encoding. In addition, the practical application of the VSN strategy is inseparable from the continuous control of the robot by employing low-level actions. [22]–[25] propose to design sim-to-real systems based on ROS for autonomous robotic navigation in the real world. In our work, we investigate the transformation from the simulated high-level discrete action space to the actual low-level continuous action space. Concretely, we work on extending the HSP learning-based VSN strategy to more realistic scenes by designing a Habitat-ROS interface.

III. Methodology

A. Problem Statement

Given a target object category G, the VSN task aims at navigating a robot to an instance of this category within a certain time T. At the beginning of each episode, the robot is spawned from a random state $s_0 \in \mathcal{S}$ in a domestic scene. At each timestep t, the VSN strategy $\pi_\theta(s_t)$ predicts the robot's action $a_t \in \mathcal{A}$ only based on the egocentric RGBD observations $o_t \in \mathcal{O}$. After executing a_t, the robot achieves the next state s_{t+1} based on the transition function $\mathcal{P}$ and receives a scalar reward $r_t \in \mathcal{R}$. In our work, the navigation task is formulated as a **M**arkov **D**ecision **P**rocess (MDP) in a DRL framework, which is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \gamma, T)$ [26]. Therefore, our DRL objective is to find a sequence of actions that maximize the total expected future rewards $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where $\gamma \in (0, 1)$ is the discount factor. Formally, the training objective of the VSN policy [27] is to minimize:

$$\begin{aligned}\mathcal{L} = ||r_t + \gamma Q_{\theta^-}(s_{t+1}, argmax_{a_{t+1}} Q_\theta(s_{t+1}, a_{t+1})) \\ - Q_\theta(s_t, a_t)||,\end{aligned} \tag{1}$$

where (s_t, a_t, r_t, s_{t+1}) is a transition uniformly sampled from an experience replay buffer. The parameters θ^- of the target network are held fixed between individual updates and updated less frequently. At each timestep, the robot gets a reward R_t consisting of four parts: (1) A time penalty term R_p that encourages the robot to choose the shortest path. (2) A collision penalty term R_c if the navigation step size is less than 0.125 m, which encourages the robot movement and punishes collision. (3) Distance reduction reward ΔD_t that denotes the robot's distance reduction to the closest object. (4) Success reward R_{succ} if the robot reaches G successfully:

$$R_t = R_p + R_c + \Delta D_t + R_{succ}, \tag{2}$$

where $R_p = -0.01$, $R_c = -0.25$ and $R_{succ} = +10$. If the robot executes the stop action when the distance from the

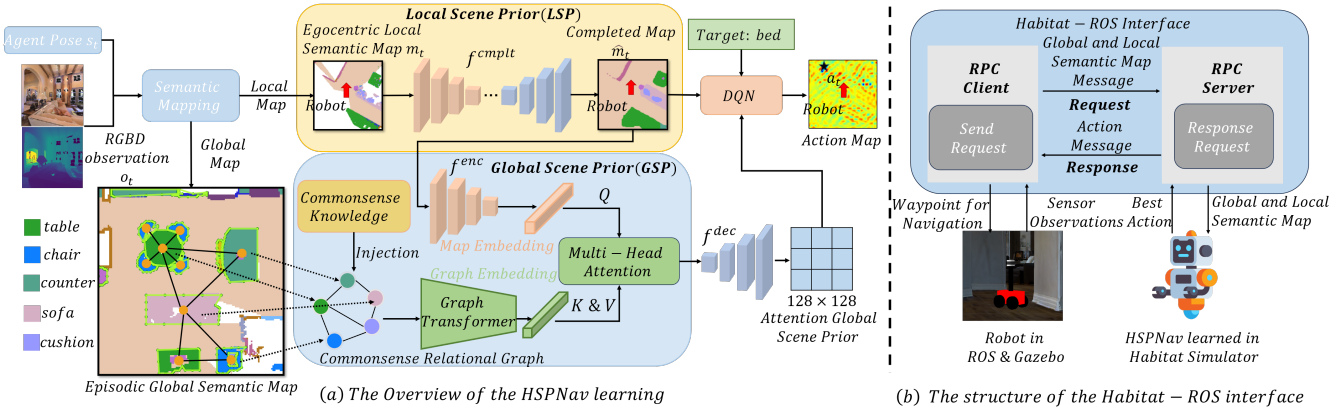

Fig. 2. (a) illustrates the overview of HSP learning. The semantic mapping module is utilized to process sophisticated sequential RGBD image frames to cumulatively build an episodic global semantic map. An egocentric local semantic map is cropped from the global map and is further semantically completed as a fine-grained LSP. A topological map is constructed as a GSP skeleton by extracting object contours and location features from the global map. Then, the commonsense knowledge embodying the subordination and co-occurrence relations among objects is injected into the skeleton to obtain the final relation graph-based GSP. The LSP is employed as a query to retrieve the scene priors from the GSP that facilitate the VSN. Finally, the object target, LSP, and retrieved global contextual memory are fed into a DQN to predict the next best navigational action. (b) illustrates the Habitat-ROS interface in a client-server construct. The client maintains a practical robot in the gazebo simulator and the server runs our VSN method HSPNav. The client sends semantic maps to the server and requests the best navigation action. The server predicts the next best action based on the HSP and feeds it back to the client.

Algorithm 1 Bayesian-based semantic fusion

```
//Each point has a semantic set consists of a semantic label set and a corresponding confidence set.
α ← 0.8
P1 ← semantic set 1, P2 ← semantic set 2
function BAYESIAN_FUSION(P1, P2)
    conf_others1 ← 1 - sum of confidences in semantic set 1
    conf_others2 ← 1 - sum of confidences in semantic set 2
    if P1 has same elements with P2 then
        Do nothing
    else
        for label in P1 not in P2 do
            Add label into P2
            conf_2(label) ← α × conf_others2
            conf_others2 ← conf_others2 − conf_2(label)
        for label in P2 not in P1 do
            Add label into P1
            conf_1(label) ← α × conf_others1
            conf_others1 ← conf_others1 − conf_1(label)
    P ← P1 × P2 with corresponding label
    return P ordered by conf
sem_set_fusion ← BAYESIAN_FUSION(P1, P2)
```

robot to G is less than 1.0 m, the VSN task is considered successful. The VSN episode terminates and fails after a fixed maximum number of timesteps. In our setting, the agent is allowed to explore up to 500 steps.

B. Hierarchical Scene Prior Learning

Most existing methods [5], [6] employ recurrent neural units to directly learn implicit NEM from raw visual observations for facilitating VSN. However, NEM is easily forgotten due to the robot's prolonged exploration and wandering. Moreover, the increase in the number of neurons required for NEM exerts pressure on the VSN strategy, which is contrary to the original intention of learning good decision-making mechanisms. In our work, we recommend the usage of semantic maps to cope with the sophisticated spatial and semantic patterns in visual observations. Empirically, humanoid agents are usually able to easily and excellently complete VSN tasks utilizing maps and landmarks without having to solidify complex visual details in their brains.

Specifically, while the agent is moving in the domestic scene, it has access to the first-view RGBD observations and corresponding camera poses. The semantic categories in RGB images are first predicted by using a pre-trained off-the-shelf segmentation module ACNet [28]. Then, the semantically segmented pixels are aligned with the depth images and further projected into 3D semantic point clouds in conjunction with the camera poses. In addition, we employ Bayesian-based semantic fusion to handle temporal-variant semantic features by fusing different point cloud frames into global point clouds. Algorithm 1 presents the semantic fusion mechanism for any two different point clouds. Finally, the global 3D semantic point clouds are projected to a top-down 2D semantic map $m_t \in R^{K \times L \times L}$. The construction and updating of global semantic maps are accompanied throughout each VSN process. $L \times L$ indicates the size of the global semantic map with $K+1$ channels. K represents the number of object categories and the extra channel indicates the unknown category.

*1) **Semantic Map-based LSP**:* Although semantics-inspired global illusions help to localize object targets, we believe a fine-grained LSP for down-to-earth navigation is indispensable. For example, assuming a person is requested to go from the living room to the kitchen to find a spoon, he should reasonably mine the semantic clues around him at present to locate his current position and the possible orientation relative to the kitchen. Inspired by this, a robot-centric local semantic map of size $M \times M$ is cropped from the global semantic map for LSP learning, as shown in Fig. 2 (a). Such local maps usually provide the robot with surrounding exhaustive semantic clues and obstacle layouts that empower robots with immersive semantic reasoning and collision avoidance capabilities. However, there are always somewhere unobserved in the local semantic map in practice

due to obstacles blocking the camera's view. The tattered semantic map only involves the partial scene priors around the robot, which may lead to wrong decisions. To tackle this issue, a **S**emantic **M**ap **C**ompletion **N**etwork (SMCN) f^{cmplt} is employed to predict the semantic clues beyond the field of view using the contextual features. The SMCN has learned the statistics of many typical room layouts after training with a vast amount of semantic maps. In particular, the SMCN takes the fragmented local semantic map m_t as input and predicts a more completed local semantic map $\hat{m}_t$.

*2) **Commonsense Relational Graph-based GSP**:* Human-like agents benefit from the commonsense knowledge accumulated over time and are thus proficient in VSN tasks. This is because humans are skilled at retrieving past experiences based on current perceptions to solve hard problems. If we regard the LSP as the robot's current immersive perceptions, the navigation experiences, i.e., the GSP, need to be further constructed. Considering there are rich subordination and co-occurrence relationships among objects in domestic scenes, our core idea is to construct a commonsense relational graph-based GSP $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ by regarding the objects as nodes $\mathcal{V}$ and the relations among them as edges $\mathcal{E}$. Specifically, the skeleton of GSP is extracted from the incrementally constructed global semantic map, while the spatial patterns are abstracted and preserved as much as possible. Then, the commonsense knowledge is injected into the skeleton to constitute the final GSP, as shown in Fig. 2 (a).

Generally, the node features $u_v \in \mathcal{V}$ consists of the semantic category embeddings ω_c, contour features $\mathcal{C} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_C, y_C)\}$, and position features (x_o, y_o). The pre-trained language model $Bert$ [29] is employed to encode each object category to get the semantic embedding $\omega_c \in \mathbb{R}^{300}$. The $MeanShift$ algorithm is utilized to cluster each object's pixels in the map to get the center position features (x_o, y_o). The contours extracting method[1] provided by $OpenCV$ is utilized to get the external contour features from the global semantic map. Considering that different objects have different textures, the contour features are exploited to distinguish object categories and preserve spatial features. The edge $e_{ij} \in \mathcal{E}$ indicates that there are spatial subordination or semantic co-occurrence relations between two object nodes v_i and v_j. All the relations are extracted from the image-captions of the **V**isual **G**enome (VG) dataset [30]. To make the relational contexts clearer, we prune and harmonize lots of object and relation aliases (for instance, "table" vs. "dining table" and "in" vs. "inside of"). It is worth noting that the weight w_{rel} of edge e_{ij} is set as the occurrence frequency of the relation in the VG dataset. Two nodes are connected with an edge only when the occurrence frequency of any relation in the VG dataset is more than 3 and the distance between them is less than 1.0 m. Therefore, each edge e_{ij} involves three-part features: (1) the spatial distance between v_i and v_j, (2) the relation embedding $\omega_r \in \mathbb{R}^{300}$, and (3) the weight of the relation w_{rel} which represents the possibility that the relation exists.

[1]https://docs.opencv.org/4.8.0/d4/d73/tutorial_py_contours_begin.html

To fully exploit the HSP, we first encode the LSP and the GSP employing a CNN encoder f^{enc} and a Graph Transformer [31] $GT(\cdot)$, respectively. Then, the LSP (map embedding) is employed as a query Q to retrieve scene priors that benefit the current VSN decision from the GSP (graph embedding), as shown in Fig. 2 (a). This process is implemented as a multi-head cross-model attention mechanism:

$$AttHead_t^i(LSP, GSP) = \frac{Q_i K_i^T}{\sqrt{d}} V_i, \tag{3}$$

$$H_t = Softmax(\underset{i\in[1,h]}{\|} AttHead_t^i(LSP, GSP)), \tag{4}$$

where $Q_i = f^{enc}(LSP)W_q^i$, $K_i = GT(GSP)W_k^i$, and $V_i = GT(GSP)W_v^i$. $\|$ means the concatenation operation and h denotes the head number. W_q^i, W_k^i, and $W_v^i \in \mathbb{R}^{d\times d}$ are learnable parameter matrices. H_t is the attention feature over the GSP.

C. Habitat-ROS Interface Towards Real Settings

The cooperation of high-performance simulators and large-scale realistic domestic scene datasets has facilitated the rapid progress of VSN. However, these simulators usually assume a high-level discrete action space that violates reality. For example, Habitat [19] employs a discrete action space of $\mathcal{A} = \{move\_forward, turn\_left, turn\_right, stop\}$ without collision avoidance. Nevertheless, the motion of robots in real scenes is continuous and requires collision avoidance. Therefore, the navigation strategies learned in the Habitat can not be directly integrated into a real robot. To tackle this challenge, we generalize our VSN policy to more realistic scenes by designing a Habitat-ROS interface.

At the communication level, the Habitat-ROS interface is a client-server construct, as shown in Fig. 2 (b). In a nutshell, the client maintains a practical robot in the gazebo[2] simulator and the server runs our HSPNav. On the client side, a scene is loaded from the MP3D [3] dataset by the gazebo before each VSN episode. Particularly, we obtain the mesh and point cloud data from the MP3D dataset, render the scene and visualize them with the gazebo and the rviz[3] plugins provided by ROS. Furthermore, a vehicle model with a continuous action space is set up in the MP3D scenes. The model is equipped with a RGBD camera and a laser sensor for environment sensing and collision avoidance, respectively. The semantic mapping techniques introduced in Subsection III-B are likewise implemented by the client based on first-view RGBD images in ROS. Through synchronous **R**emote **P**rocedure **C**all (RPC) style communication services provided by ROS, the client transports the completed local map $\hat{m}_t$ and the global map to the server.

The server receives the maps and responds to the client by feeding back the next best action. To achieve the transition from discrete actions to continuous actions, HSPNav infers the best navigation action $a_{t+1}^* = (x_p^t, y_p^t)$ as a coordinate on the local map, as shown in Fig. 2 (a). The client converts the

[2]https://gazebosim.org/home

[3]http://wiki.ros.org/rviz

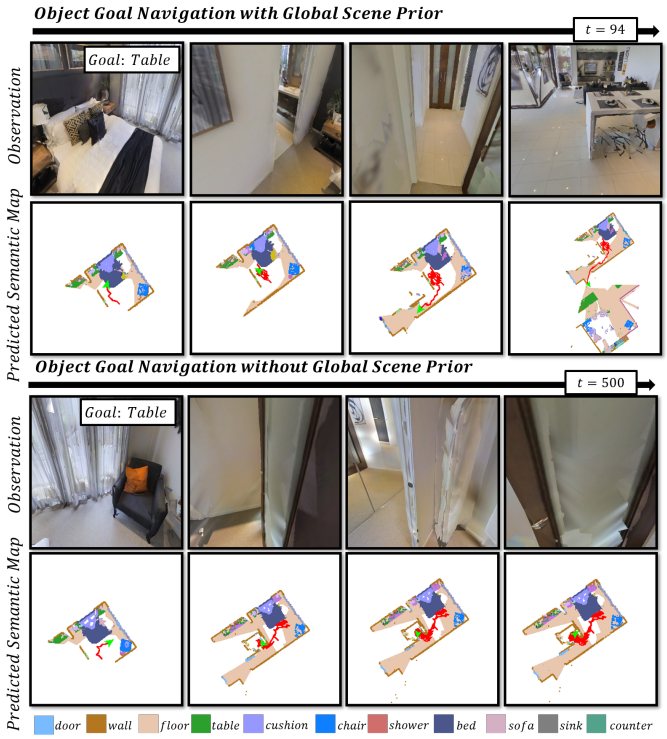

Fig. 3. illustrates the qualitative results of VSN with and without GSP attendance. The robot is requested to find and navigate to a table. It is obvious that the attendance of the GSP facilitates rescuing the robot from its dilemma and navigating it to the correct object instance.

received best action into a real-world navigational sub-goal with the following transformations:

$$\beta = arctan\frac{|y_p^t - y_c^t|}{|x_p^t - x_c^t|}, \tag{5}$$

$$x_w^{t+1} = x_w^t + rcos(\beta), y_w^{t+1} = y_w^t + rsin(\beta), \tag{6}$$

where (x_c^t, y_c^t) is the center of $\hat{m}_t$ and (x_w, y_w) is the robot's current position in gazebo simulator. r = 0.25 m is the navigation step size in our practical setting. Once the navigational sub-goal is specified, the classical navigation pipeline can be employed to flexibly and continuously navigate the robot to the corresponding coordinate while avoiding collisions. Compared to frequent pauses in discrete mode, the robot can navigate to the target object smoothly, stopping only at a few sub-goals.

IV. Experiments

A. Experimental Setup and Implementation Details

1) ***Experimental Setup****:* In this section, we first conduct comparative and ablative studies using the Habitat simulator and the standard train-test split of the MP3D dataset. Then, the proposed Habitat-ROS interface is employed to generalize the HSPNav strategy to more realistic scenes with continuous action spaces. The size and resolution of the robot's action space $\mathcal{A}$ (i.e., the action map) are set to 128×128 and 0.05 m, respectively. The target object categories of the training episodes are uniformly sampled from the domestic scenes. 687 test VSN episodes are generated from the test split of the MP3D with the target object in $\{Bed, Counter, Shower, Sink, Sofa, Table, Toilet\}$. It is worth noting that all training is performed on the train split, and the test split is previously unknown to the robot.

TABLE I

Statistical comparative experimental results.

Method	SR(%) ↑	SPL ↑
SAVN [6]	0.9	0.009
SemExp [7]	25.8	0.118
ZSON [17]	22.9	0.067
HSPNav	**28.1**	**0.135**

For comparative and ablative studies, navigation **S**uccess **R**ate (SR) and **S**uccess weighted by **P**ath **L**ength (SPL) [32] are employed as evaluation metrics. The SR metric measures the effectiveness of the VSN strategy and is formulated as $\frac{1}{N}\sum_{i=1}^{N} S_i$, where N stands for the total number of the test episodes and S_i is the binary success indicator of i-th episode. The SPL metric evaluates the efficiency of the VSN strategy that is given by $\frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{max(l_i,p_i)}$, where p_i denotes the path length generated by the robot in episode i and l_i is the length of the shortest path to the closest target object instance in episode i. Our proposed method is compared against the following 3 baselines: SAVN [6], SemExp [7], and ZSON [17].

2) ***Implementation Details****:* The ACNet model is employed to semantically segment RGB images to construct global and local semantic maps. To avoid domain gaps, the ACNet model is retrained with 209,200 RGBD images of 40 object categories from the train split of the MP3D dataset. The SMCN is implemented with an encoder-decoder architecture that has 4 down-sample residual blocks and 5 up-sample residual blocks. Moreover, the SMCN has learned the statistics of many typical room layouts after training with more than 50,000 ground-truth top-down maps. Note that all learnable parameter training is performed on the train split to avoid data leakage. We employed one NVIDIA 3090 GPU for rendering the simulation scenes and training the models. We train our approach utilizing the train split of the MP3D dataset over 200k steps. The discount factor γ for DRL is set to 0.99. The network parameters are trained by the Adam optimizer with a learning rate of 0.0001. The threshold for Bayesian semantic fusion in Algorithm 1 is set to α = 0.8.

B. Comparative Studies

By comparing our method with 3 baselines, we report the experimental results in Table I. Our method outperforms all baselines in terms of the SR and SPL metrics. SemExp extracts semantic priors by constructing global semantic maps and samples a middle-level navigation target to decide where to look for objects. Statistically, our method relatively improves 8.91% and 14.41% in terms of SR and SPL compared with SemExp, respectively. Such improvements demonstrate that our HSP is more conducive to VSN than purely learning semantic priors from episodic global semantic maps. Unlike SemExp, SAVN learns semantic priors and navigational skills directly from raw visual observations in an end-to-end manner. However, SAVN performs poorly in both SR and SPL. As justified in [1], the reason is that directly processing sophisticated visual images may lead to poor

TABLE II

STATISTICAL ABLATION EXPERIMENTAL RESULTS (SR(%) / SPL). -GSP: ONLY USE THE LSP. +GT: REPLACE ACNET'S PREDICTIONS USING GROUND-TRUTH SEMANTIC SEGMENTATION.

Method	bed	counter	shower	sink	sofa	table	toilet	Avg
HSPNav-GSP(SSCNav [8])	4.2/0.040	10.6/**0.022**	7.1/0.009	**23.2/0.076**	5.7/0.029	38.8/**0.261**	3.6/0.033	24.6/**0.148**
HSPNav	**14.6/0.086**	**13.6**/0.021	**17.1/0.018**	18.2/0.034	**7.5/0.043**	**42.9**/0.235	**3.6/0.033**	**28.1**/0.135
HSPNav-GSP+GT(SSCNav [8])	10.4/0.057	22.7/0.092	25.7/0.073	28.0/0.149	24.5/0.115	65.6/0.425	14.3/**0.061**	43.8/0.259
HSPNav+GT	**33.3/0.223**	**57.6/0.325**	**61.4/0.195**	**54.9/0.278**	**83.0/0.690**	**89.7/0.481**	**21.4**/0.056	**72.3/0.481**

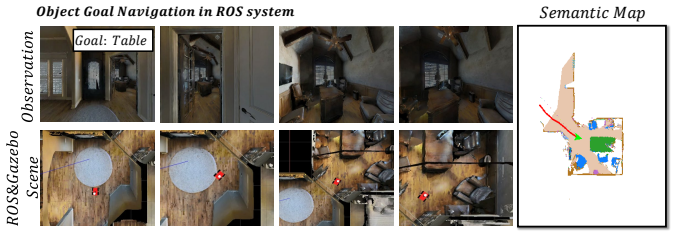

Fig. 4. illustrates the qualitative results of VSN using the Habitat-ROS interface. The robot moves in a more realistic gazebo scene with a continuous action space. Both the HSP learning and the navigational reasoning are performed on the server.

generalization performance and robustness. Notably, SAVN is developed based on the AI2THOR [4] dataset. There are concealed differences between AI2THOR-based and MP3D-based image rendering in terms of texture, lighting, and so on. These differences are fatal for VSN strategies that strongly rely on raw visual observation.

Compared to ZSON, our method relatively improves 22.71% and 101.49% in terms of SR and SPL. Such results indicate that our hierarchical scene priors outperform the visual-language experiences provided by CLIP. Benefiting from our HSP learning technique and experience query mechanism, HSPNav achieves promising navigation performance. In other words, we exploit the scene priors more reasonably by incorporating spatial and semantic features such as fine-grained map features, relationships among objects, and object contour features into the VSN policy learning.

C. Ablations for Global Scene Prior

The contributions of mapless, global semantic map-based, and visual-language scene priors to VSN have been adequately compared in Subsection IV-B. In this subsection, we focus on ablating the global semantic prior and studying its impact on HSPNav. Table II reports the VSN performance of HSPNav with and without GSP (-GSP) for different object categories. SSCNav [8] utilizes an SMCN to complete the egocentric local semantic map and only exploits the local scene priors to infer the next action. By treating SSCNav as an ablation study of the GSP (HSPNav-GSP), we find that the attendance of GSP (HSPNav vs. HSPNav-GSP) improves the SR metric for almost all object categories. Intuitively, the average SR metric is relatively improved by 14.23%. In addition, using the GSP also results in competitive SPL metrics. To investigate the effect of semantic segmentation accuracy on HSPNav, we further replace the predicted output of ACNet with ground-truth semantic segmentation which does not cause navigation failures due to semantic segmentation error. As expected, the usage of ground-truth semantic segmentation (HSPNav vs. HSPNav+GT) greatly improves the performance of HSPNav, achieving an average SR metric of 72.3%. Moreover, even without the GSP, the navigation performance of the HSPNav-GSP+GT is significantly improved. On the one hand, such experimental results demonstrate that HSPNav is sensitive to semantic segmentation accuracy. On the other hand, they also indicate the outstanding contribution of our GSP to VSN. Therefore, the improvement in semantic segmentation accuracy will lead to a significant increase in the performance of HSPNav. In addition, we find that the usage of GSP leads to more wandering of the robot in simple VSN tasks compared to HSPNav-GSP, which leads to slightly lower SPL metrics but higher SR metrics. However, the global semantic inspiration provided by GSP helps to improve the performance of difficult VSN tasks. For example, Fig. 3 illustrates the qualitative results of VSN with and without GSP. We find the attendance of the GSP facilitates rescuing the robot from its dilemma and navigating it to the correct object instance.

D. Qualitative Evaluation for Habitat-ROS Interface

To verify the practicality of our Habitat-ROS interface, a client is constructed based on the scene numbered 8WUmhLawc2A comes from the MP3D dataset, as shown in Fig. 4. As an example, the robot is randomly initialized in the scene and requested to find a table. The client constructs global and local semantic maps using the robot's visual observations and sends them to the server to request appropriate navigation sub-goals. As expected, the feedback received by the robot triggers the classic navigation pipeline. The robot is continuously navigated to the specified object instance successfully. Similarly, our Habitat-ROS interface allows HSPNav to process requests from the robot in the real world, requiring only the alignment of map data formats.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a VSN strategy HSPNav based on HSP learning towards the previously unknown scenes. The completed fine-grained local semantic map is treated as an LSP to retrieve scene priors from the GSP that are conducive to the VSN. Sufficient comparative and ablative studies demonstrate the superiority of our method and the effectiveness of HSP. As expected, we find that the increase in semantic segmentation accuracy will drastically improve the performance of HSPNav, which reflects the great potential of our approach. In addition, we work on extending HSPNav to more practical scenes with continuous action spaces by designing a Habitat-ROS interface. The qualitative evaluation validates the feasibility of our idea. In future work, we will attempt to deploy HSPNav to a real robot platform and further improve its generalization capabilities to diverse scenes and unseen object categories.

REFERENCES

[1] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023.

[2] 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.

[3] 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)*, IEEE, 2017.

[4] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, *et al.*, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[5] J. Ye, D. Batra, A. Das, and E. Wijmans, "Auxiliary tasks and exploration enable objectgoal navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16117–16126, 2021.

[6] M. Wortsman, K. Ehsani, M. Rastegari, A. Farhadi, and R. Mottaghi, "Learning to learn how to learn: Self-adaptive visual navigation using meta-learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6750–6759, 2019.

[7] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[8] Y. Liang, B. Chen, and S. Song, "Sscnav: Confidence-aware semantic scene completion for visual semantic navigation," in *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 13194–13200, IEEE, 2021.

[9] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*, IEEE, 2022.

[10] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] G. Georgakis, B. Bucher, K. Schmeckpeper, S. Singh, and K. Daniilidis, "Learning to map for active semantic goal navigation," in *International Conference on Learning Representations (ICLR)*, 2022.

[12] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao, "Bevbert: Topo-metric map pre-training for language-guided navigation," *arXiv preprint arXiv:2212.04385*, 2022.

[13] B. Chen, J. Kang, P. Zhong, Y. Cui, S. Lu, Y. Liang, and J. Wang, "Think holistically, act down-to-earth: A semantic navigation strategy with continuous environmental representation and multi-step forward planning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[14] B. Chen, S. Lu, P. Zhong, Y. Cui, Y. Liang, and J. Wang, "Semnav-hro: A target-driven semantic navigation strategy with human–robot–object ternary fusion," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107370, 2024.

[15] K. Fang, A. Toshev, L. Fei-Fei, and S. Savarese, "Scene memory transformer for embodied agents in long-horizon tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 538–547, 2019.

[16] M. Fortunato, M. Tan, R. Faulkner, S. Hansen, A. Puigdomènech Badia, G. Buttimore, C. Deck, J. Z. Leibo, and C. Blundell, "Generalization of reinforcement learners with working and episodic memory," *Advances in neural information processing systems*, vol. 32, 2019.

[17] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "Zson: Zero-shot object-goal navigation using multimodal goal embeddings," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[19] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

[20] H. Bharadhwaj, Z. Wang, Y. Bengio, and L. Paull, "A data-efficient framework for training and sim-to-real transfer of navigation policies," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 782–788, IEEE, 2019.

[21] P. Anderson, A. Shrivastava, J. Truong, A. Majumdar, D. Parikh, D. Batra, and S. Lee, "Sim-to-real transfer for vision-and-language navigation," in *Conference on Robot Learning*, pp. 671–681, PMLR, 2021.

[22] Q. Zou, Q. Sun, L. Chen, B. Nie, and Q. Li, "A comparative analysis of lidar slam-based indoor navigation for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6907–6921, 2021.

[23] H. Hu, K. Zhang, A. H. Tan, M. Ruan, C. Agia, and G. Nejat, "A sim-to-real pipeline for deep reinforcement learning for autonomous robot navigation in cluttered rough terrain," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6569–6576, 2021.

[24] B. Chen, P. Zhong, Y. Cui, S. Lu, Y. Liang, and Y. Sheng, "Emexplorer: an episodic memory enhanced autonomous exploration strategy with voronoi domain conversion and invalid action masking," *Complex & Intelligent Systems*, pp. 1–15, 2023.

[25] B. Chen, Y. Cui, P. Zhong, W. yang, Y. Liang, and J. Wang, "Stexplorer: A hierarchical autonomous exploration strategy with spatio-temporal awareness for aerial robots," *ACM Transactions on Intelligent Systems and Technology*.

[26] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3357–3364, IEEE, 2017.

[27] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[28] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1440–1444, 2019.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[31] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, "Masked label prediction: Unified message passing model for semi-supervised classification," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 1548–1554, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[32] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.