Graph Attention for Spatial Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Imbuing robots with human-levels of intelligence is a longstanding goal of AI research. A critical aspect of human-level intelligence is spatial reasoning. Spatial reasoning requires a robot to reason about relationships among objects in an environment to estimate the positions of unseen objects. In this work, we introduced a novel graph attention approach for predicting the locations of query objects in partially observable environments. We found that our approach achieved state of the art results on object location prediction tasks. Then, we evaluated our approach on never before seen objects, and we observed zero-shot generalization to estimate the positions of new object types.

1 Introduction

The goal of discrete spatial reasoning is to predict the likelihood that a query object exists at a location from observations with associated locations (1). The problem is critical to autonomous navigation under uncertainty (2). A key difference from other types of inference is that the observation data are not independently and identically distributed (IID), because of the location information. In this work, we focus on an application to robot navigation. Figure 1 illustrates contextual reasoning for discrete spatial prediction. For example, observing the position of a refrigerator and counter provides strong evidence for the position of the stove.

Embodied visual navigation requires an agent to use visual perception to control navigation through an environment (3). Recent progress has been made in this field with the introduction of embodied environments both scanned(4; 5) and simulated(6; 7). Several research challenges have been introduced that require a robot to navigate to the location of unseen objects(8; 2). It is assumed that the agent has no prior knowledge of the room layout or contents. A query object category is given to the agent. The challenges require embodied agents to develop a semantic understanding of the composition and spatial associations between objects either explicitly or implicitly.

In this work, we introduce an approach called Graph Attention for Spatial Prediction (*GRASP*). We hypothesized that the spatial relationships among observed objects provide strong priors that could be used to estimate the position of unobserved query objects. Our query-based representation of these spatial relationships relies on (i) an allocentric graph representation of spatially situated objects in the environment, (ii) query-based attention over the graph, and (iii) dimensionality-reduced object embeddings. Our results quantify empirically two key advantages of our approach. *GRASP* is able to reason over diverse layouts, object types, and quantities of objects, and our approach generalizes to new object types that were not observed during training.

2 Related Work

Prior work has explored visual navigation through embodied reasoning challenges including ImageGoal (9; 10), RoomGoal (11; 12), ObjectNav (13; 2), and PointGoal (11; 12; 14). Instruction-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

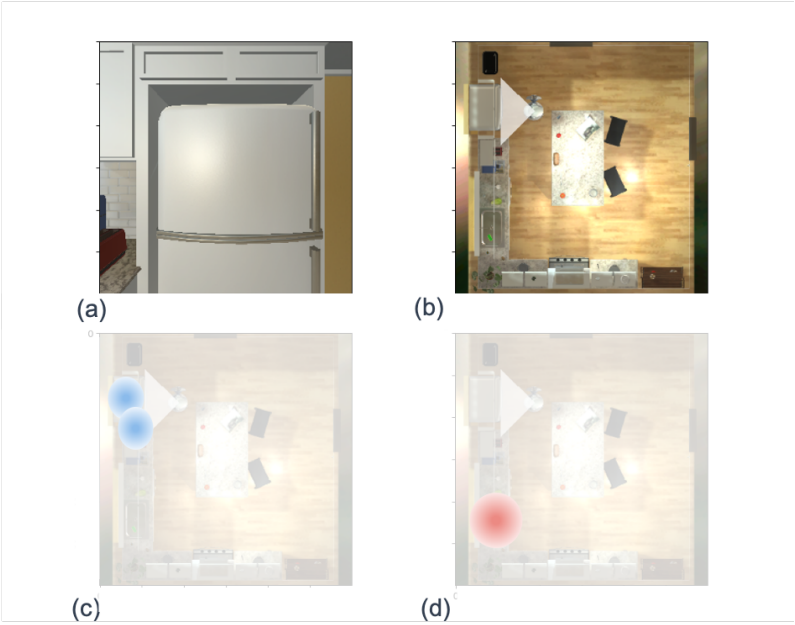

Figure 1: An illustration of our approach to embodied contextual reasoning for spatial prediction. In this example, the query object is the sink. (a) Egocentric robot view of the AI2Thor environment. Several items including the refrigerator, counter, and cabinets were observed (b) The observations are maintained in an situated allocentric view of the environment. (c) Allocentric attention (blue) is used to identify observations with relevance to the the query object. (d) Inferred position (red) of the query object relative to the robot

following challenges require agents to follow a sequence of instructions (8). Progress on these challenges has been made with research into topological graphs (9; 15), recurrent networks (16; 17; 14), and allocentric spatial memory (18; 19; 20). Our work explores new ways of estimating spatial positions of unseen objects which is relevant for object navigation tasks.

Spatial relationship features can provide additional contextual and semantic relationships (21). The use of these features was explored using tree-based approaches (22; 23). Spatial data from multiple sources can be fused into a common reference to provide additional context for spatial predictions (24; 25).

ObjectNav tasks have object types as the navigation goal. End-to-end reinforcement learning methods have improved state-of-the-art with improved state representations (26; 27), data augmentation (28), and auxiliary tasks (29) improve object navigation performance to new scenes. Examples of improved state representations include the use of visual attention (26) and state priors (27).

Modular RL methods for object navigation separate the task into object localization and point navigation (30; 31; 32). These approaches use offline pretraining to learn high likelihood regions for object localization. Then, point navigation is used to move the robot towards those high likelihood regions. Most recently, state of the art performance was achieved by learning potential functions to highlight exploration frontiers with a high chance of resulting in the object being found (32). In this work, we focus on the object localization task.

Imitation learning has been used to find object navigation policies from demonstrations (33; 34; 35) and fine-tuning (34; 35). Self-supervised approaches have been explored that learn distance and semantic label scores from image collections (15).

3 Problem Definition

Object localization is a kind of spatial prediction that involves predicting the likelihood of a query object type q at position y from observations of other discrete objects and their explanatory features.

Formally, let the environment be represented by $\mathbf{X} = \{x(\mathbf{s_i}) | i \in \mathcal{N}, 1 \leq i \leq n\}$ where $\mathcal{N}$ is the set of observed objects of size n, $\mathbf{s_i}$ is the coordinate of the object relative to an origin point $\mathbf{s_0}$, and $x(\mathbf{s_i})$ is the feature vector of explanatory variables associated with the object at position $\mathbf{s_i}$. The feature vector consists of an object embedding and position information described in more detail in Section 4. For a given set of observed objects $\mathbf{X}$ and query or target object q, we aim to estimate (i) the position $\hat{y}$ of the closest instance of the query object and (b) the occupancy grid $\hat{z}$ for the query object relative to the agent's position. To evaluate these estimates, we define two metrics:

Closest Point Metric – Relative to an agent's position defined to be the origin $\mathbf{s_0}$, The closest instance of the query object $\hat{y}$ is represented as $\hat{y} = \mathcal{F}(\mathbf{X}, q) = (r, sin(\theta), cos(\theta))$ where (r, θ) are polar coordinate representation relative to $\mathbf{s_0}$. θ is represented by $sin(\theta)$ and $cos(\theta)$ to avoid discontinuities at 2π. The absence of the object is indicated by r being greater than a threshold t. We define the closest point metric as the L2-norm between the ground truth y and $\hat{y}$.

Occupancy Map Metric – Estimates of the query-specific occupancy map $\hat{z}$ are represented as a matrix $\mathbf{M}$ where $\mathbf{M}_{i,j}$ is the probability of the query at position i, j relative to the agent's position. We define the occupancy map metric as the l2-norm between the ground truth occupancy map z and estimates of the occupancy map $\hat{z}$.

4 Graph Attention for Spatial Prediction

In the following section, we introduce *GRASP* in more detail. Figure 1 illustrates our approach to spatial reasoning based on allocentric graph attention.

To define the feature vector $x(\mathbf{s_i})$ for each object, we include the vector of position $\mathbf{s_i}$ relative to $\mathbf{s_0}$ as polar coordinates and a pretrained word embedding $\mathcal{W}$ of the named type of object i.

$$x(\mathbf{s_i}) = [\mathbf{s_i} - \mathbf{s_0}, \mathcal{W}(\ell_i)]$$

where ℓ_i is the name of the type of the object at $\mathbf{s_i}$. In our experiments, we used the Numberbatch embedding (36) reduced to the first 32 output elements as $\mathcal{W}$.

We represent observations as a graph $G(V, E)$. A vertex v_i is defined for each observed object i as an attention-weighted composite embedding of the relevant object. Specifically, the vertex features make use of a learned embedding function $\mathcal{E}$ and a learned saliency function $\mathcal{S}$. Let $\mathcal{S}(x(\mathbf{s_i}), q)$ estimate the probability that the object at position $\mathbf{s_i}$ is relevant for estimating the closest instance of object type q.

$$V(\mathbf{X}, q) = [\mathcal{S}(x(\mathbf{s_i}), q)\mathcal{E}(x(\mathbf{s_i})), 1 \leq i \leq n]$$

An edge $e_{i,j} \in E$ exists if $||\mathbf{s_i} - \mathbf{s_j}||_2 < \delta$. In our experiments, we set δ to 3 meters.

The output of the inference model is then

$$\mathcal{F}(V, q, E) = \text{GNN}(V(\mathbf{X}, q), E(\mathbf{X}), q).$$

which produces an estimate of the closest instance of the query object in modified polar coordinates (r, $sin(\theta)$, $cos(\theta)$) where θ is then infered from $sin(\theta)$ and $cos(\theta)$.

The functions $\mathcal{E}$ and $\mathcal{S}$ are approximated as MLP deep networks with parameters η and ν respectively. The function F is approximated as a convolutional graph neural network followed by MLP layers with parameters γ.

Network inference produces closest instance estimates relative to the agent's position in polar coordinates (r, θ) represented as $(r, sin(\theta), cos(\theta))$ to avoid discontinuities in θ.

5 Results

We organized experiments to evaluate (i) how well *GRASP* predicts the location of the closest instance and (ii) the occupancy grid associated with the object type of interest, and (iii) the ability of *GRASP* to generalize to objects not seen during training .

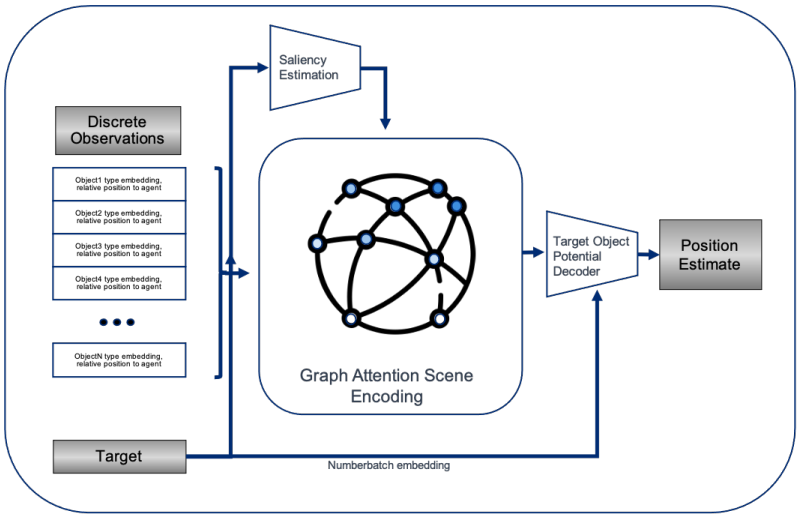

Figure 2: Overview of the graph attention approach to egocentric scene embedding. Variable sized list of discrete object observations are encoded as nodes in a graph where edges encode distances between objects. Node representations are reweighted by attention to the target object. The decoder estimates the position estimate for the closest instance of the target object.

Dataset – The AI2Thor (6) illustrated in Figure 1 is a benchmark environment for indoor navigation and task completion. iThor contains 120 indoor layouts and 2000 distinct object types. Public competitions such as the Alfred challenge (8) make use of AI2thor to evaluate autonomous task completion algorithms. We make use of AI2thor to both train and evaluate several approaches to spatial prediction.

Training – We collected a dataset consisting of tuples $(\mathbf{s_0}, \mathbf{X}, q, y, z)$ where $\mathbf{s_0}$ is the agent's position, $\mathbf{X}$ is the list of observed objects and positions, q is the query object type, y is the agent relative position of the closest instance of the query object type, and z is the occupancy map centered at the agent's position as defined in Section 3. To ensure similar data distributions for all approaches, the same 80/20 split of layouts was used for training and evaluation of each approach. Training was conducted on 1 million batches of gradient descent using the adam optimizer over 96 room layouts and 14882 dataset tuples. We trained all approaches for 67 epochs.

Metrics – For each method, we measure both the closest point metric and the occupancy grid metric. To facilitate comparison between methods that estimate either $\hat{z}$ or $\hat{y}$, we introduced an approach to estimate one from the other. Given $\hat{z}$, $\hat{y}$ is estimated as the position with maximum probability. Likewise $\hat{y}$ can be used to estimate $\hat{z}$ with a Gaussian probability distribution centered at $\hat{y}$.

Evaluation – All results that we report in Table 1 are the result of averaging over 1000 samples from the test dataset. Samples from the test set are derived from 24 held-out room layouts. Each test sample was derived from a random selection of test layout, agent position, query object type, and observed objects.

Comparisons and Baselines – In this section, we describe both strong and weak baselines that we compared in our experiments. We included a *Uniform Grid* baseline that assigns equal probability over all positions in the occupancy grid. *Random Grid* assigns random probability over all positions in the occupancy grid. *PONI [(32)]* is a recent state-of-the-art approach for object location prediction. The approach makes use of a *UNet* (37) architecture to highlight exploration frontiers within an occupancy grid with high likelihood of containing the object. In addition to PONI, we also tested another variant of the UNet architecture with Gaussian blur to better account for positional uncertainty. These approaches make use of a one hot embedding over channels for object type representations. The training loss for these approaches is MSE between predicted and ground truth grid distance. PONI makes (32) use of dense environment representations. Dense environment representations use of occupancy grids to represent the positions of observed objects. We included several approaches with dense representations for comparison. To encode object types, these methods made use of one-hot encoding across channels. *VAE [(38)]* is a conditional variational autoencoder for predicting

the occupancy grid for the query object location. We tested a variant of VAE with Gaussian blur applied to the output to account for positional uncertainty. These approaches make use of a one-hot embedding over channels for object type representations. The training loss for these approaches is MSE between predicted and ground truth grid distance along with the typical KL loss term for controlling the latent distribution.

			Novel Layouts		Novel Layouts and Objects	
Approach	**Backbone**	**Representation**	**Grid Distance**	**Point Distance**	**Grid Distance**	**Point Distance**
uniform grid	baseline	NA	3.95	9.41	3.15	10
random grid	baseline	NA	3.96	9.39	3.15	10
PONI	UNET	Dense	2.67	9.54	3.15	10
UNET+blur	UNET	Dense	2.59	9.47	3.15	10
VAE	VAE	Dense	2.89	9.5	3.16	10
VAE+blur	VAE	Dense	2.09	9.53	3.16	10
GRASP (Ours)	GNN	Sparse	**1.79**	**2.02**	**1.91**	**2.34**

Table 1: Summary of results for two experimental settings: (i) novel room layouts, and (ii) novel room layouts and novel query objects. We reported occupancy grid distance and closest instance point distance for each approach. Approaches are organized by network type and observation encoding. Additional method details include environment representation and neural network backbone when applicable.

5.1 Generalizing to New Layouts

In this experiment, we evaluated the ability of the approaches described in section 5 to generalize to new layouts. Training and evaluation were conducted as described in Sections 5 and 5. The results of our experiments are summarized in Table 1. We found that *GRASP* out performed the other baselines and state of the art approaches like PONI (32). The sparse allocentric graph representation of observations in the environment is a key difference between *GRASP* and the others in the comparison. We found that approaches that relied on dense representations of the environment like PONI, VAE, and UNet-blur performed better than chance for grid distance prediction but struggled to perform over chance for closest point prediction. The difference may be due to the conversion from an occupancy map output to a point prediction. The point estimate was taken to be the centroid of the region with highest probability. While the centroid may be close to a real instance of the query object, it may not be the closest instance to the robots position.

5.2 Generalizing to New Layouts and New Object Types

In our second experiment, we evaluated the ability of the approaches described in Section 5 to generalize to new layout and object types that were not observed during training. During training, we withheld 10% of object types at random from the dataset. During evaluation, query objects were only selected from the set of withheld object types. Other than these changes, training and evaluation were conducted as described in Sections 5 and 5.

The results of the experiment are described in Table 1. We found that *GRASP* out performed baseline and state-of-the-art approaches like PONI (32). A key design difference between *GRASP* and the others was the use of dimensionality-reduced object embeddings for representations. In this study, we used pretrained and frozen Numberbatch (36) embeddings. This embedding has the property that the representation of semantically similar objects are close in the output space. The semantic relationships between objects withheld from training and those included in training allowed *GRASP* to successfully make correct predictions where the other approaches failed.

6 Conclusions

In this work, we introduced a new approach for object location prediction called *GRASP*. We demonstrated empirically that our approach achieves state of the art performance for predicting the location of unobserved objects based on the occupancy grid and closest instance metrics. Our experiments demonstrated that *GRASP* was able generalize to new layouts and object types without retraining.

References

[1] M. B. Smyth and J. Webster, *Discrete Spatial Models*. Dordrecht: Springer Netherlands, 2007, pp. 713–798. [Online]. Available: https://doi.org/10.1007/978-1-4020-5587-4_12

[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] 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.

[4] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," *arXiv preprint arXiv:2109.08238*, 2021.

[5] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[6] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[7] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9068–9079.

[8] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 740–10 749.

[9] N. Savinov, A. Dosovitskiy, and V. Koltun, "Semi-parametric topological memory for navigation," *arXiv preprint arXiv:1803.00653*, 2018.

[10] 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)*. IEEE, 2017, pp. 3357–3364.

[11] 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.

[12] M. Savva, A. X. Chang, A. Dosovitskiy, T. Funkhouser, and V. Koltun, "Minos: Multimodal indoor simulator for navigation in complex environments," *arXiv preprint arXiv:1712.03931*, 2017.

[13] 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.

[14] 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*, 2019, pp. 9339–9347.

[15] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 875–12 884.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[18] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," *arXiv preprint arXiv:1903.01959*, 2019.

[19] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2616–2625.

[20] J. F. Henriques and A. Vedaldi, "Mapnet: An allocentric spatial memory for mapping environments," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8476–8484.

[21] R. Frank, M. Ester, and A. Knobbe, "A multi-relational approach to spatial classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 309–318.

[22] A. McGovern, N. Troutman, R. A. Brown, J. K. Williams, and J. Abernethy, "Enhanced spatiotemporal relational probability trees and forests," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 398–433, 2013.

[23] A. McGovern, N. C. Hiers, M. Collier, D. J. Gagne II, and R. A. Brown, "Spatiotemporal relational probability trees: An introduction," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 935–940.

[24] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1436–1444.

[25] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang, "Semantic annotation of mobility data using social media," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1253–1263.

[26] B. Mayo, T. Hazan, and A. Tal, "Visual navigation with spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 898–16 907.

[27] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," *arXiv preprint arXiv:1810.06543*, 2018.

[28] O. Maksymets, V. Cartillier, A. Gokaslan, E. Wijmans, W. Galuba, S. Lee, and D. Batra, "Thda: Treasure hunt data augmentation for semantic navigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 374–15 383.

[29] 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*, 2021, pp. 16 117–16 126.

[30] 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.

[31] 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)*. IEEE, 2021, pp. 13 194–13 200.

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[33] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.

[34] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," *arXiv preprint arXiv:1903.01959*, 2019.

[35] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1–10.

[36] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[38] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.