RSMPNet: Relationship Guided Semantic Map Prediction

Jingwen Sun, Jing Wu, Ze Ji, Yu-Kun Lai
Cardiff University, Cardiff, UK
{sunj39, wuj11, jiz1, laiy4}@cardiff.ac.uk

Abstract

In semantic navigation, a top-down map with accurate and complete semantic information is vital to subsequent decision-making. However, due to occlusions and limitations of the robot's field of view (FOV), there are often unobserved areas in the top-down maps. To address this problem, recent works have studied semantic map prediction to complete the top-down maps. In this work, we propose to improve map prediction by integrating relational information. We propose RSMPNet, a relationship-guided semantic map prediction network, which makes use of semantic and spatial relationships to predict unobserved areas from accumulated semantic maps. Specifically, we propose a Relationship Reasoning Layer that includes two modules, namely 1) the Semantic Relationship Graph Reasoning Module (SeGRM) to capture the semantic relationship and 2) the Spatial Relationship Graph Reasoning Module (SpGRM) to utilize the spatial relationship. We also design a semantic relationship enhanced loss to enhance our model to learn semantic relationship information. Experiments show the effectiveness of our proposed network which achieves state-of-the-art performance in semantic map prediction. Our code and dataset are publicly available at https://github.com/jws39/semantic-map-prediction

1. Introduction

Indoor robots operating autonomously in unstructured real-world environments are increasingly required to help with people's life and work these years. Rather than navigating from one point to another using a pre-built metric map, these "intelligent" robots are expected to be able to navigate to an object or a room in an unseen environment according to some given semantic information. This kind of task is defined as semantic navigation and has been increasingly studied recently. In semantic navigation, the robot should navigate to the target with the capability of inferring or reasoning the semantic layout of the environment based on current observations. Specifically, a top-down map containing the objects' layout in a 3D environment can facilitate subsequent decision-making and planning to improve the robot's performance [8, 30]. However, most works [8, 9, 11, 21, 25, 30] only build the semantic map in the robot's field of view (FOV). In such a way, the robot can only get local and incomplete information, due to the limited FOV and the occlusions that often occur in complex indoor environments.

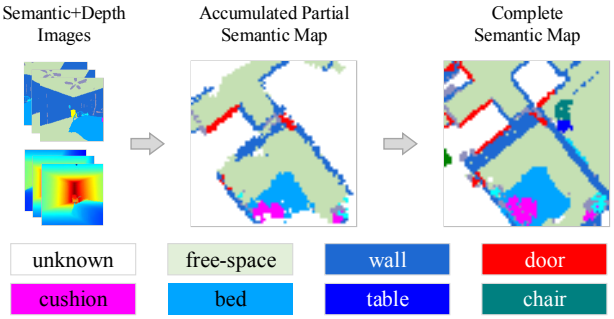


Figure 1. Semantic Map Prediction. The robot takes semantic and depth images as input and then predicts the complete map from the accumulated partial map that it has seen.

To solve these problems, researchers [6, 12, 13, 17, 20, 23, 27, 29] have started to study map prediction, which is also the aim of our work. As shown in Fig. 1, while the robot navigates in the environment, it takes semantic (obtained from some semantic segmentation algorithms) and depth images as input and then predicts the complete map from the accumulated partial map that it has seen. For the map prediction task, some previous works [12, 13, 20, 23, 27, 29] model geometric information only, like occupancy and layout, while other works [6, 17] infer semantics in unknown areas from the partial map to help the robot navigate to an object. However, these works usually make use of the information of object categories only without considering the relationships between them. We observe that the relationships between objects play an important role in human navigation in a new environment. We can infer what is in the

This work of Sun is supported in part by the China Scholarship Council (202006120013).

unobserved area using this prior knowledge. For example, a sink is usually in the kitchen, and chairs tend to be near a table. So such relationship information should also be considered in semantic map prediction.

Based on these observations, we propose a **R**elationship-guided **S**emantic **M**ap **P**rediction **Net**work (RSMPNet), a network that predicts the semantic map from observations using semantic and spatial relationships. The semantic relationship measures how similar two objects are in their semantic meanings. For example, pens and pencils are semantically similar. The spatial relationship measures how close two objects are in their spatial locations. For example, chairs are usually found around a table. To model semantic relationships, we utilize the similarity between categories at every pair of pixels in the semantic map. In addition, to model spatial relationships, for each pixel, we calculate its distances to all other pixels to build the prior knowledge such that nearby pixels contribute more to a pixel's representation. Accordingly, motivated by Guan *et al*. [7], we propose a **Relationship Reasoning Layer** that explores the integration of a **Se**mantic Relationship **G**raph **R**easoning **M**odule (SeGRM) and a **Sp**atial Relationship **G**raph **R**easoning **M**odule (SpGRM) to learn prior semantic and spatial knowledge to enhance the feature representation effectively for map prediction. And we also design a new loss function to enhance the ability to learn the semantic relationship. Our contributions are as follows:

- We consider both semantic and spatial information in map prediction and evaluate different designs of a Relationship Reasoning Layer to capture semantic and spatial relationships.
- We also design a semantic relationship enhanced loss to improve the learning ability of the semantic relationship, which considers the prediction accuracy of both the map and the relationship.
- Extensive experiments show that our method can effectively learn semantic and spatial relationships to improve the performance of semantic map prediction.

2. Related Work

2.1. Map Prediction

Some researchers [12, 13, 20, 23, 27, 29] have focused on predicting unobserved regions from partial observations to improve the robot's navigation performance. These works only predict an occupancy map using metric information to decide the subsequent goal. Kapil *et al*. [13] propose a method to predict the occupancy map of the area beyond the current FOV using a deep generative network. And they also take into account the uncertainty and ambiguity in mapping and exploration by using a multi-head network to get multiple predictions. Recently, some researchers [6, 17] have studied semantic map prediction. Liang *et al*. [17] first collect local top-down semantic maps with randomly removed regions and then use scene completion to predict unobserved regions. Georgakis *et al*. [6] first predict an occupancy map and then combine it with the single view top-down semantic map to get the full semantic map. These two works are related to our work, but they only use individual categorical information and do not consider using relationship information for map prediction. Instead, our focus is to extract more useful relationship information to improve semantic map prediction.

2.2. Semantic Navigation

With the rise of interactive simulator platforms [15, 26, 33], end-to-end learning-based approaches [9, 11, 19, 21, 25, 28, 35, 37] to semantic navigation have been widely studied. Some works [19, 28, 35, 37] directly use images as input and do not require a memory module, like a map. In such mapless navigation, the observations are encoded with networks directly, and the encoded representations are sent to the policy network, together with the embedding of the target, to generate an action. While for map-based navigation, an explicit semantic map with multi-channel one-hot-like representation is used to represent the environment's semantic information (e.g. objects and rooms) in some works [1, 3]. Some other works [8, 9, 11, 21, 25, 30] use implicit neural representations that encode semantic information for effective action decision-making. However, these works do not consider the relationships among objects when building the map for navigation.

2.3. Relationship Guided Navigation

Rather than directly using semantic maps that only contain object classes, some works [4, 5, 18, 22, 31, 34] consider the relationships among objects for semantic navigation as well. Yang *et al*. [34] propose to use a Graph Convolutional Network (GCN) [14] for incorporating prior knowledge that encodes the object relationship extracted from the Visual Genome [16] dataset into a deep reinforcement learning framework. The robot uses the features from the knowledge graph for predicting corresponding actions. Some other works [5, 18, 22] also use GCNs to encode the relationship information, and the difference between them is the definition of the nodes. Different from the above works, which construct a graph to encode the relationship information, Druon *et al*. [4] use a context grid to represent this kind of semantic information. These works consider the object relationship during navigation, while in our work, we focus on leveraging the semantic and spatial relationships for enhanced semantic map prediction.

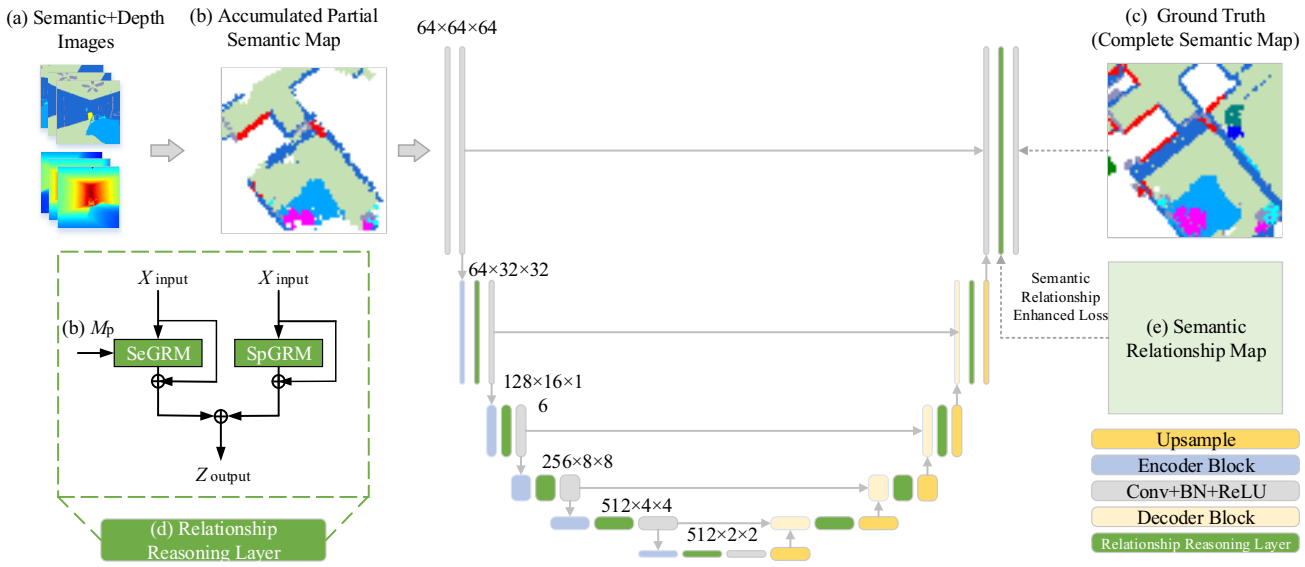


Figure 2. RSMPNet Overview. The input to the network is an accumulated partial semantic map (b) by aggregating the projected semantic maps from previous semantic and depth images (a). The output is the expected semantic map prediction (c). The Relationship Reasoning Layer (d) in the green box consists of SeGRM and SpGRM, which capture semantic and spatial relationships, respectively. The Semantic Relationship Map (e) is the supervision of SeGRM in the last Relationship Reasoning Layer to enhance the capability of the network to learn semantic relationships. (Notation: M_p Accumulated Partial Semantic Map, $\oplus$ add operation.)

3. Our Method

The overall network architecture is shown in Fig. 2. The input to the network is an accumulated partial semantic map, which is obtained by aggregating the projected semantic maps from previous observations (semantic and depth images) as in the work [6]. The output is expected to be the complete semantic map. The network architecture is based on UNet [24], and the Relationship Reasoning Layer is inserted after each encoder and decoder layer. The Relationship Reasoning Layer consists of Semantic Graph Reasoning Module (SeGRM) which captures the semantic relationship, and Spatial Graph Reasoning Module (SpGRM) which captures the spatial relationship. Both SeGRM and SpGRM are based on graph convolutional neural networks (GCNs) [14]. The Semantic Relationship Map is used in semantic relationship enhanced loss to make the network better learn the semantic relationship. In the following subsections, we will first briefly introduce GCNs, and then describe the Relationship Reasoning Layer and the Semantic Relationship Enhanced Loss in detail.

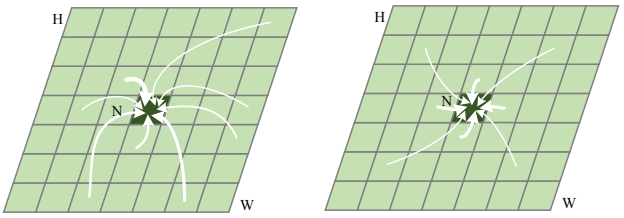
3.1. Graph Convolutional Networks

Graph Convolutional Networks (GCNs) generalize the convolution operation from grid data to graph structures [32]. A graph is represented as $G = (V, E)$, where V is the set of nodes, and E is the set of edges. The main idea is to update the representation of one node by aggregating its own features and neighbors' features. The operation in one layer of a GCN can be expressed as:

$$Z = \sigma(\hat{A} \cdot X \cdot \Theta), \tag{1}$$

where X, Z are the input and output features. Θ is the learnable parameter of the layer in GCN. $\hat{A}$ is the adjacency matrix representing relationships among nodes. $\sigma(\cdot)$ denotes a non-linear activation function.

GCN is widely used for capturing relational information. Some works [12, 13, 20, 23, 27, 29] have used GCNs to extract semantic relationships from a pre-built object graph. In these works, a node usually denotes an object, while in our method, nodes are pixels, and edges are the semantic or spatial relationships among pixels. Fig. 3 is an illustration of our two graph reasoning modules using GCNs to capture the semantic and spatial relationships. In both modules, the feature of node N is updated from the features of other nodes. The weights of different nodes are represented by the thickness of the lines with the arrow. For the semantic relationship, some far nodes can still have big weights because of their similar semantic meanings, while for the spatial relationship, only nearby nodes have big weights because of their close distances.



(a) SeGRM (b) SpGRM

Figure 3. Illustration of Relationship Graph Reasoning Modules. In SeGRM, the node N is updated according to the distance to other nodes in semantic meaning. In SpGRM, the node N is updated according to the distance to other nodes in spatial terms. The weights of different nodes are represented by the thickness of the lines with the arrow.

3.2. Relationship Reasoning Layer

The Relationship Reasoning Layer consists of SeGRM and SpGRM. These two modules are designed following the work of Guan *et al*. [7]. However, instead of aggregating the two modules sequentially as in [7], we aggregate these two modules in a parallel way, which has demonstrated better performance in our experiments. Below we give a brief description of the two modules.

The SeGRM module aims to enhance the features with the semantic relationship. GCN [14] is used to capture the semantic relationship and update the features. As shown in Fig. 3 (a), every pixel is a node in the graph, and edges encode the semantic relationship of the observed area, represented by the semantic-aware adjacency matrix $\hat{A}_{Se}$. Suppose $X \in \mathbb{R}^{H\times W\times D}$ is the original feature map after an encoder layer (with width W, height H and number of channels D), which contains high-level information. C is the number of object categories. W_1 and W_2 are two learnable parameter vectors to extract the high-level relationship $\hat{A}_{Obj} \in \mathbb{R}^{C\times C}$ as in Eq. (2). To combine this kind of high-level information and also consider low-level information, the input partial semantic map, $M_p \in \mathbb{R}^{H\times W\times C}$, and $\hat{A}_{Obj}$ are utilized to compute the semantic relationship, $\hat{A}_{Se} \in \mathbb{R}^{HW\times HW}$, as in Eq. (3). Then through SeGRM, the feature map is iteratively updated using $\hat{A}_{Se}$ as in Eq. (1).

$$\hat{A}_{Obj} = W_1 X W_2,\ \hat{A}_{Obj} \in \mathbb{R}^{C\times C}, \tag{2}$$

$$\hat{A}_{Se} = M_p \hat{A}_{Obj} M_p^T,\ M_p \in \mathbb{R}^{H\times W\times C},\ \hat{A}_{Se} \in \mathbb{R}^{HW\times HW}, \tag{3}$$

The SpGRM module aims to enhance features by aggregating information based on the spatial relationship. GCN is also used to do the propagation with Eq. (1). Accordingly, the spatial-aware adjacency matrix, $\hat{A}_{Sp} \in \mathbb{R}^{HW\times HW}$, is designed as below to capture such information:

$$\hat{A}_{Sp}(x_i, x_j) = distance(x_i, x_j),\ \hat{A}_{Sp} \in \mathbb{R}^{HW\times HW}, \tag{4}$$

where $distance(\cdot,\cdot)$, following [7], is the Manhattan distance between the two pixel locations x_i and x_j. Then the feature of each pixel is updated using $\hat{A}_{Sp}$ through the GCN.

The Relationship Reasoning Layer aggregates the SeGRM module and the SpGRM module. In the work of Guan *et al*. [7], they aggregate these two modules sequentially. However, sequentially aggregating the two types of relationships may result in the latter relationship overriding the former one in the output feature maps. Therefore, in our work, we aggregate these two modules parallelly. We first send the features to SeGRM and SpGRM modules individually, and then add the outputs of these two modules together (as shown in Fig. 2 (d)). The parallel aggregation does show better results than the sequential aggregation in our experiments.

3.3. Semantic Relationship Enhanced Loss

To enhance the ability to learn the semantic relationship, we design the loss function considering not only the accuracy of the map prediction, but also the accuracy of the semantic relationship. The loss function is as follows:

$$L = \lambda_M L_M + (1-\lambda_M) L_A, \tag{5}$$

where L_M, L_A denote the primary map prediction loss and the semantic relationship loss. L_M is the cross-entropy of the predicted and the ground-truth semantic maps. For L_A, following the work [36], we first convert the given semantic map to a one-hot encoding $G \in \mathbb{R}^{H\times W\times C}$, and then calculate the ground-truth semantic relationship map as $A_{Se} = GG^T$. The semantic relationship loss is calculated as the binary cross-entropy between the semantic-aware adjacency matrix at the last relationship reasoning layer and the ground-truth semantic relationship map. λ_M is the weight to balance prediction and semantic relationship loss. We empirically set $\lambda_M = 0.3$.

4. Experiments

In this section, we carry out extensive experiments to validate our proposed method. We first introduce how we collect the dataset. Next, we present the implementation details. And then, we demonstrate the effectiveness of the proposed modules in the ablation study and compare the results with the state-of-the-art map prediction methods.

4.1. Data Collection

The aim of our work is to predict the unseen areas in a top-down semantic map from the areas that have been seen.

Method	Encoder	Decoder	SeGRM	SpGRM	Sequ	Para	Sem Loss	mIoU↑	mF1↑	mAcc↑
M0	ResNet18	✓						24.87	0.3845	33.18
M1	ResNet18	✓	✓					25.79	0.3959	33.83
M2	ResNet18	✓		✓				25.97	0.3987	34.20
M3	ResNet18	✓	✓				✓	26.23	0.4025	34.66
M4	ResNet18	✓	✓	✓	✓		✓	26.11	0.4008	34.52
M5	ResNet18	✓	✓	✓		✓	✓	**26.33**	**0.4048**	**35.21**

Table 1. Ablation Study. We add modules one by one to evaluate their effectiveness. A ✓indicates that the corresponding module is added. **SeGRM/SpGRM**: the Semantic/Spatial part in Relationship Reasoning Module. **Sequ/Para**: Combining these two modules in sequential/parallel way. **Sem Loss**: Semantic Relationship Enhance Loss.

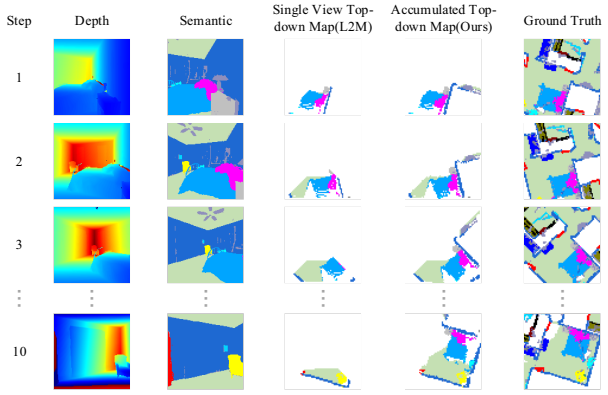


Figure 4. Semantic Map Prediction Data. Columns 3 and 4 show the single-view top-down map in L2M [6] and the accumulated top-down map in our method.

Thus, the input to the neural network is a partial semantic map that has been accumulated from previous observations, and the output is a predicted complete semantic map. To allow supervised training of the neural network, a large set of such partial semantic maps with known ground-truth complete semantic maps are required. A dataset with a similar purpose has been collected in L2M [6]. However, the semantic partial maps in that dataset are from the current single observation only without accumulating previous observations. We think that in the application of navigation, accumulated semantic maps are more reasonable, as they contain more information and can be easily built with depth images. In SSCNav [17], accumulated maps are collected. However, they remove some regions randomly, which may not be realistic in real-world applications.

Therefore, following the work in [6], in Fig. 4, we show an example of data collection for an episode with 10 steps. We directly use depth images and the semantic images given by the simulator to compute the top-down map. First, we project the 2D images into 3D space with semantic information in semantic images to get points. The ground floor plane is divided into a 64×64 grid. In each grid point, we get the frequency of every semantic label and compute a probability distribution over the semantic labels. The category of each grid point is set to the object with the maximum probability. When the next partial semantic map is obtained, it is aggregated to the previous accumulated partial map according to the pose changed by the agent. The overlapping area is updated by multiplying the per-category probabilities, which are then normalized to sum to 1 in every grid. The robot is in the center of the top-down map. The columns 3 and 4 show the single view top-down map in L2M [6] and the accumulated top-down map in our dataset. The top-down map is 64×64 resolution, and every pixel denotes 0.1 m in reality. We generate a dataset with 39256, 5100, and 5404 accumulated top-down maps as the training, validation, and test sets from Matterport3D (MP3D) [2] dataset using the Habitat [26] simulator. In our dataset, we chose 27 common objects from 41 categories of objects in the MP3D dataset, as in the work [6]. As shown in Fig. 5, we count the frequency and area ratio of each object in our dataset.

4.2. Implementation Details

We use a pre-trained ResNet-18 [10] to initialize the encoder backbone. We train the whole model using the Adam optimizer with a learning rate of 0.0002 and a batch size of 8. The training is carried out on a single NVIDIA TESLA V100 GPU and takes about 20 hours for 50 epochs. In addition, we adopt the mean Intersection over Union (mIoU), mean F1-measure (mF1) and mean pixel accuracy (mAcc) as the evaluation metrics.

4.3. Ablation Study

We first carry out experiments to demonstrate the effectiveness of each component in our method (i.e., the SeGRM, the SpGRM, and the Semantic Relationship Enhanced Loss), and then compare the two ways of aggregation mentioned in Section 3.2 (sequential vs. parallel).

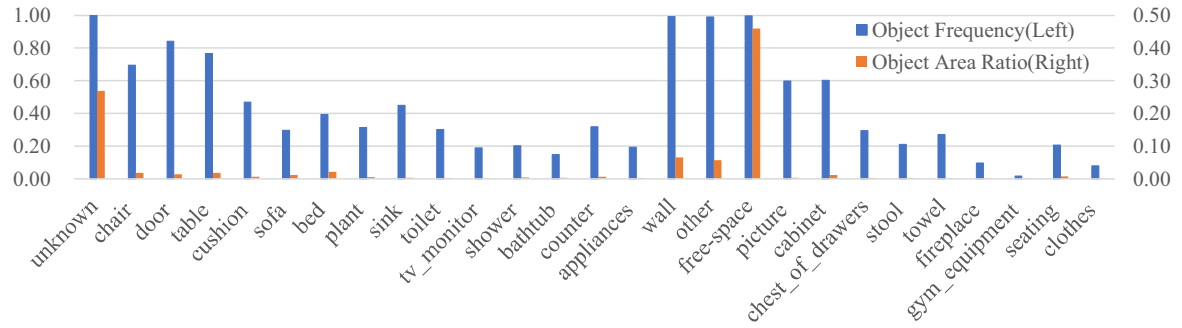


Figure 5. Dataset Statistics. We count the frequency and area ratio of each object in our dataset.

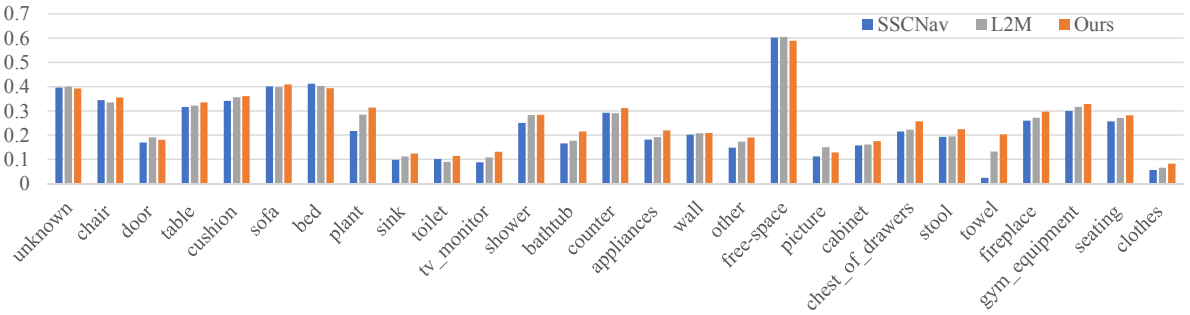


Figure 6. Semantic Map Prediction Results (mIoU). Our method has better results in 22 out of 27 categories than the other two methods in mIoU.

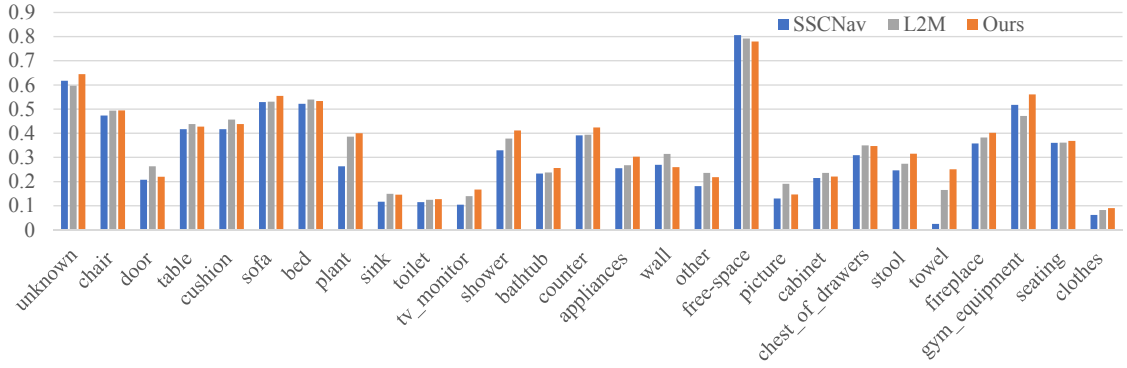


Figure 7. Semantic Map Prediction Results (mAcc). Our method has better results in 16 out of 27 categories than the other two methods in mAcc.

Table 1 shows the quantitative results.

We use the model in L2M [6] as our baseline (see M0 in Table 2), which is a UNet model with five encoder and decoder convolutional blocks and skip connections. We first add the SeGRM module after each encoder and decoder block (see M1 in Table 1). We can see an absolute improvement of 0.92% on mIoU, 0.011 on mF1, and 0.65% on mAcc, indicating the importance of semantic relationship to semantic map prediction. We then evaluate the performance by only adding the SpGRM module after each encoder and decoder block (see M2 in Table 1). We can see that the SpGRM module improves the mIoU, mF1, and mAcc by 1.1%, 0.014, and 1.02% over the baseline model. The improvements demonstrate the effectiveness of the SpGRM

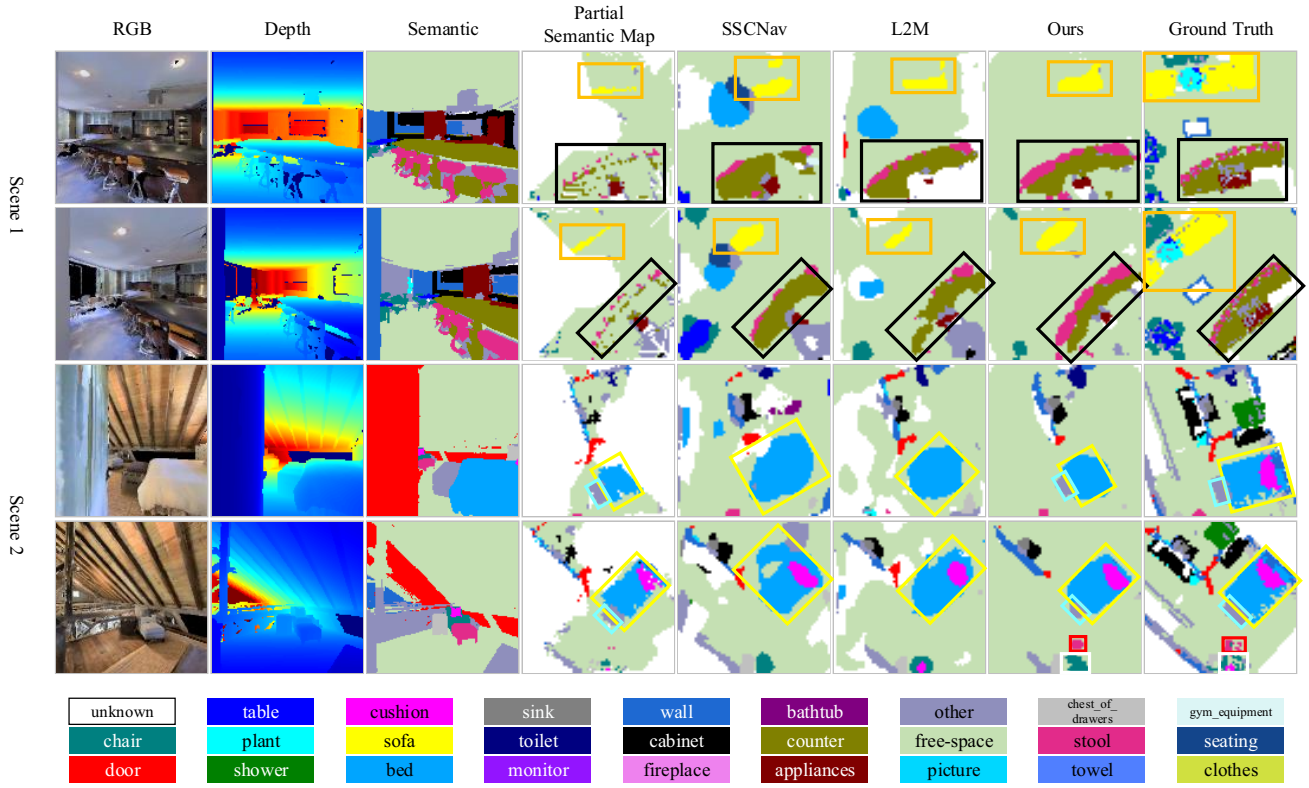


Figure 8. Qualitative Comparison. We compare our method, SSCNav [17] and L2M [6] different methods on Scene 1 (8194nk5LbLH) and Scene 2 (EU6Fwq7SyZv). The results show that our method can not only use semantic and spatial relationships to predict objects in unobserved areas (rows 1, 2 and 4), but also retain observed objects better (rows 3 and 4).

module. To evaluate the effectiveness of the Semantic Relationship Enhanced Loss (see M3 in Table 1), we add the loss to the last SeGRM module based on M1. The results obtain 26.23% and 34.66% in terms of mIoU and mAcc, surpassing the M1 by 0.44% and 0.83%. This result shows that the loss can further enhance the performance. We also conduct experiments to explore how these two modules are aggregated (see M4 and M5 in Table 1). The results show that parallel aggregation achieves better results than sequential aggregation.

We also perform experiments with different values of λ_M. As shown in Table 2, we set $\lambda_M = [0.1, 0.3, 0.5, 0.7, 0.9]$. The results show $\lambda_M = 0.3$ is the best value for our network.

4.4. Comparison Study

For semantic map prediction, there are currently two similar works, SSCNav [17] and L2M [6]. These two works use the dataset built by themselves to train their models. For comparison, we train and evaluate their models on our dataset. Table 3 shows the quantitative results. The first row is the result of No Prediction, which directly uses the accumulated partial map to calculate the metrics. Compared with other methods, we can see that map prediction can effectively predict unobserved areas and improve the accuracy of the map. The second row is the result of the model used in SSCNav, a simple encoder-decoder architecture based on ResNet. The third row is our baseline, the UNet model used in L2M. From the last three rows, we can see that our method achieves 26.33% mIoU, 0.4048 mF1, and 35.21% mAcc, outperforming the other two methods on all the metrics. We also show the prediction results of

λ_M	mIoU↑	mF1↑	mAcc↑
0.1	25.95	0.3988	34.45
0.3	**26.33**	**0.4048**	**35.21**
0.5	26.13	0.4005	34.87
0.7	26.24	0.4028	34.92
0.9	25.87	0.3983	34.64

Table 2. Ablation study about λ_M . We do some experiments with different values of λ_M.

Method	mIoU↑	mF1↑	mAcc↑
No Prediction	16.69	0.2777	21.48
SSCNav [17]	23.35	0.3621	31.37
L2M [6]	24.87	0.3845	33.18
Ours	**26.33**	**0.4048**	**35.21**

Table 3. Comparison on our dataset. We compare our method with SSCNav [17] and L2M [6] and also show the No Prediction results.

different methods on the 27 objects in Fig. 6 and Fig. 7. Our method has better results in 22 and 16 out of 27 categories than the other two in mIoU and mAcc, respectively.

4.5. Qualitative Results

In this section, we evaluate our method qualitatively, as shown in Fig. 8. We compare our method, SSCNav [17] and L2M [6] on the same episode to show the prediction results when the robot navigates in a new environment.

From the results, we can observe that SSCNav [17] predicts some wrong objects, such as the bed and free-space classes, while the prediction of L2M [6] is better than SSCNav. We think this is attributed to the skip connection in UNet used in L2M compared to a simple encoder-decoder network used in SSCNav.

We notice our method has better prediction in areas close to the observed areas. For example, the stools around the counter are better predicted (the black boxes in rows 1 and 2) in Scene 1. We believe this is attributed to the spatial relationship that better aggregates information in nearby areas. Moreover, our method can preserve the observed area better than the other methods. For example, the bed and other classes (the yellow and cyan boxes in rows 3 and 4) are better preserved than the other two methods in Scene 2.

We also notice the sofa (the orange boxes in Scene 1) is better predicted compared to the other two methods in Scene 1. Although the sofa is spatially far from the other objects, it is semantically close to the stools. The semantic relationship used in our method thus mutually helped with the prediction of both the stools and the sofa. In Scene 2, the stool and the chair (the red and white boxes in row 4) are close both in spatial locations and semantic meanings. So these two objects can mutually improve each other's predictions. On the contrary, without the constraints of the semantic relationship, there are more wrong predictions of irrelevant objects using the other two methods, such as the prediction of bed (blue). We also show the prediction results on an episode in Fig. 9. Our method can predict the unobserved area and reserve the observed area well.

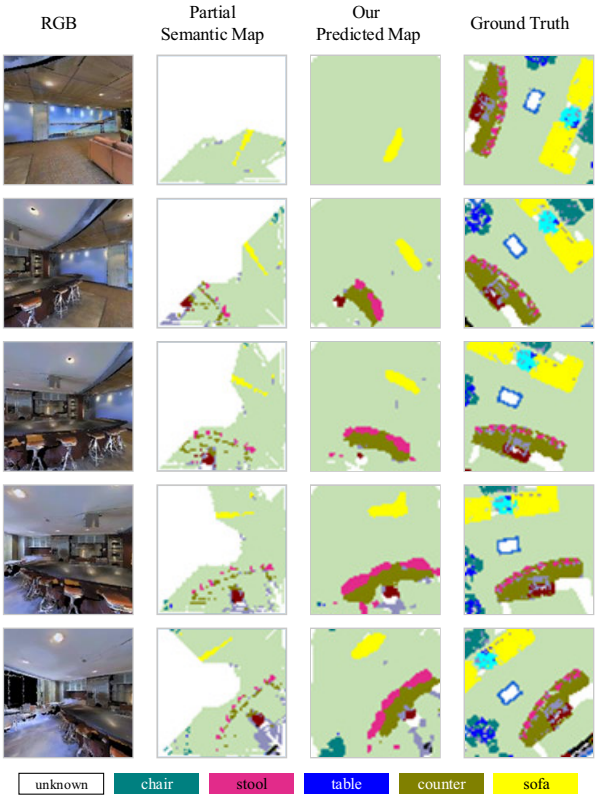


Figure 9. Prediction Results on An Episode. We show the results of our method on an episode every two steps.

4.6. Limitations

The semantic and spatial relationships can improve the prediction of relevant objects. However, it also restricts the prediction of other less relevant objects. This makes our method behave more conservatively in prediction compared to the other two methods. How to balance the pros and cons needs further investigation. An immediate direction of future work is to integrate the relationship reasoning module into different encoder/decoder layers adaptively.

5. Conclusions

In this paper, we propose a Relationship Reasoning Layer including two modules, SeGRM and SpGRM, to learn semantic and spatial relationships to improve the performance of semantic map prediction. We also design a loss function to enhance the learning of the semantic relationship and explore how to aggregate the SeGRM and the SpGRM modules. Experiments show that our method can outperform the state-of-the-art map prediction methods.

References

[1] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic MapNet: Building allocentric semantic maps and representations from egocentric views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 964–972, 2021. 2

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[3] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 2

[4] Raphael Druon, Yusuke Yoshiyasu, Asako Kanezaki, and Alassane Watt. Visual object search by learning spatial context. *IEEE Robotics and Automation Letters*, 5(2):1279–1286, 2020. 2

[5] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *European Conference on Computer Vision*, pages 19–34. Springer, 2020. 2

[6] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. *arXiv preprint arXiv:2106.15648*, 2021. 1, 2, 3, 5, 6, 7, 8

[7] Huankang Guan, Jiaying Lin, and Rynson WH Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2022. 2, 4

[8] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2017. 1, 2

[9] Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. Unifying map and landmark based representations for visual navigation. *arXiv preprint arXiv:1712.08125*, 2017. 1, 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[11] Joao F Henriques and Andrea Vedaldi. MapNet: An allocentric spatial memory for mapping environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8476–8484, 2018. 1, 2

[12] Yuki Katsumata, Akinori Kanechika, Akira Taniguchi, Lotfi El Hafi, Yoshinobu Hagiwara, and Tadahiro Taniguchi. Map completion from partial observation using the global structure of multiple environmental maps. *arXiv preprint arXiv:2103.09071*, 2021. 1, 2, 3

[13] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *International Conference on Robotics and Automation (ICRA)*, pages 5453–5459. IEEE, 2019. 1, 2, 3

[14] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 3, 4

[15] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 2

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 2

[17] Yiqing Liang, Boyuan Chen, and Shuran Song. SSC-Nav: Confidence-aware semantic scene completion for visual semantic navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 13194–13200. IEEE, 2021. 1, 2, 5, 7, 8

[18] Yunlian Lyu, Yimin Shi, and Xianggang Zhang. Improving target-driven visual navigation with attention on 3D spatial relationships. *Neural Processing Letters*, pages 1–20, 2022. 2

[19] Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8846–8852. IEEE, 2019. 2

[20] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. In *European Conference on Computer Vision*, pages 513–529. Springer, 2020. 1, 2, 3

[21] Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. *arXiv preprint arXiv:1702.08360*, 2017. 1, 2

[22] Yiding Qiu, Anwesan Pal, and Henrik I Christensen. Learning hierarchical relationships for object-goal navigation. *arXiv preprint arXiv:2003.06749*, 2020. 2

[23] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *European Conference on Computer Vision*, pages 400–418. Springer, 2020. 1, 2, 3

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[25] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018. 1, 2

[26] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE*

International Conference on Computer Vision, pages 9339–9347, 2019. 2, 5

[27] Vishnu Dutt Sharma, Jingxi Chen, Abhinav Shrivastava, and Pratap Tokekar. Occupancy map prediction for improved indoor robot navigation. *arXiv preprint arXiv:2203.04177*, 2022. 1, 2, 3

[28] William B Shen, Danfei Xu, Yuke Zhu, Leonidas J Guibas, Li Fei-Fei, and Silvio Savarese. Situational fusion of visual representation for visual navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2881–2890, 2019. 2

[29] Rakesh Shrestha, Fei-Peng Tian, Wei Feng, Ping Tan, and Richard Vaughan. Learned map prediction for enhanced mobile robot exploration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1197–1204. IEEE, 2019. 1, 2, 3

[30] Saim Wani, Shivansh Patel, Unnat Jain, Angel Chang, and Manolis Savva. MultiON: Benchmarking semantic map memory using multi-object navigation. *Advances in Neural Information Processing Systems*, 33:9700–9712, 2020. 1, 2

[31] Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, and Yuandong Tian. Bayesian relational memory for semantic visual navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2769–2779, 2019. 2

[32] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020. 3

[33] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 2

[34] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018. 2

[35] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectnav. *arXiv preprint arXiv:2104.04112*, 2021. 2

[36] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12416–12425, 2020. 4

[37] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3357–3364. IEEE, 2017. 2