SMeLt: 2D Semantic Map Prediction with Latent Context

Wang Jiaming (A0040976Y)
Jamie-w@nus.edu.sg
School of Computing, NUS

Introduction

The utilization of a 2D semantic map is prevalent in object goal navigation tasks. SOTA techniques employ a pipeline (fig. 1) to construct the semantic map, which involves the following steps:

(1) Generation of a 3D point cloud,
(2) Object detection using pretrained models
(3) Projection of the 3D point cloud onto a 2D grid map

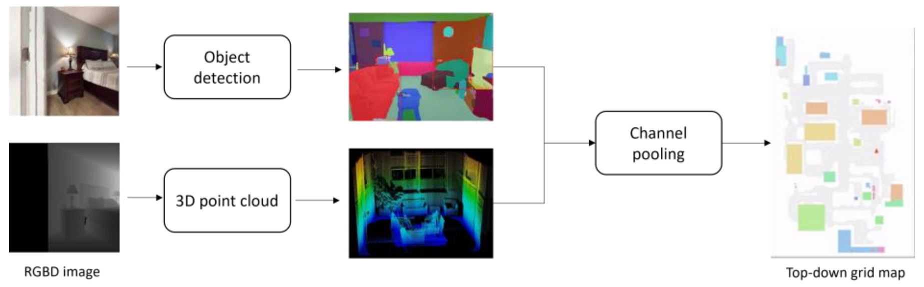

Fig. 1 Classic pipeline to construct the semantic map

However, detecting small objects poses challenges to this approach due to noisy sensors and low resolution. We propose the **S**emantic **M**ap Prediction with **L**atent Context (SMeLt) framework to tackle the challenges associated with detecting small objects during the 2D semantic mapping procedure, by leveraging contextual information around the object (fig. 2)

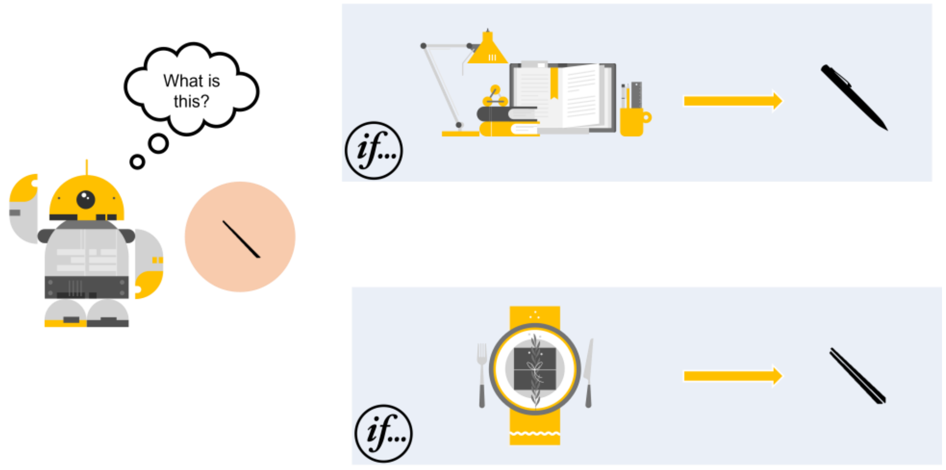

Fig. 2 Human can effectively use context information in the object recognition process.

Our main contributions are:

- The introduction of a novel framework that uses contextual information to improve the semantic mapping accuracy;
- Analysis of the proposed method;
- The presentation of a challenging dataset that focuses on small object detection and mapping task.

Methodology

Task definition

The task is to generate a 2D grid map, represented by a tensor of dimensions $C \times H \times W$, from the RGBD observation of an agent. Each entry in the tensor corresponds to a physical cell in the world that denotes the presence of a specific object within in that cell. C represents the total number of distinct object classes.

Formulation

We formulate the problem as a a simplified probabilistic graphical model (PGM) as depicted in fig. 3. Within this model, the latent context is represented by z, the semantic label s is a binary random variable, and he RGBD observation is denoted by o.

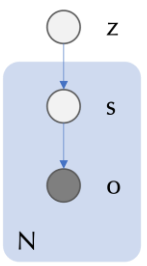

Fig. 3 PGM

$$\begin{aligned}
\mathcal{L} &= \sum_n log\, p(s,o) \\
&= \sum_n log(\sum_z p(s|o,z)p(o)p(z)) \\
&\propto \sum_n log(\mathbb{E}_z(p(s|o,z))) \\
&\geq \sum_n \mathbb{E}_z\, log\, p(s|o,z) \\
&= \sum_n \mathbb{E}_{z\sim q(z)} \frac{p(z)}{q(z)} log\, p(s|o,z)
\end{aligned}$$

Given the above PGM, we can derive our training objective by maximizing the observed data log likelihood $\mathcal{L}$ as shown in above derivation.

SMeLt framework

We amortize $q(z)$ as a neural network f_θ that learns to encode the contextual information into some latent variable z. We parametrize $p(s|o,z)$ as a visual encoder f_ϕ, and a map decoder f_ω, as illustrated in fig. 4

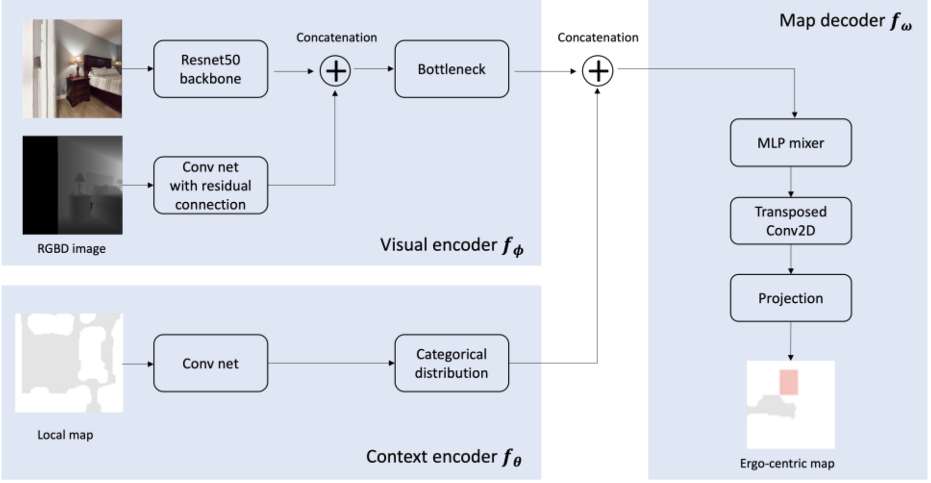

Fig. 4 Overview of the SMeLT framework

The neutral networks can be trained using the following algorithm:

```
Given $f_\theta$: $q(z)$; $f_\phi$ and $f_\omega$: $p(s|o,z)$; m: number of samples per observation;

1. Sample $z^m \sim q(z)$ for m=1,…M
2. Approximate $\sum_n \mathbb{E}_{z\sim q(z)} \frac{p(z)}{q(z)} log\, p(s|o,z) \approx \sum_n \frac{1}{m} \sum_m \frac{p(z^m)}{q(z^m)} log\, p(s|o,z^m)$
3. Calculate weight cross entropy loss $\mathcal{L} = \sum_n \frac{1}{m} \sum_m \frac{p(z^m)}{q(z^m)} (f_\theta(o,z^m) logs + (1 - f_\theta(o,z^m)) log(1-s))$
4. Update $\phi, \theta$ with $\frac{\partial \mathcal{L}}{\partial \phi}, \frac{\partial \mathcal{L}}{\partial \theta}$
```

Dataset

We rendered 127,732 RGBD images with ground truth semantic map from randomly sampled view-points in the Matterport 3D semantics dataset using Habitat simulator. We have published this dataset on Kaggle to foster further research on this task.

Experiments

We compared our proposed method with point cloud reconstruction pipeline (PC), and a baseline system that does not utilize the contextual information (table 1). We observe a 34% and 11% improvement when using the contextual information, measured by cross-entropy (CE) error on predicted semantic map for all object classes and small object classes respectively.

Model	CE error (all objects)	CE error (small objects)
PC	0.291	failed
W/o context z:	0.055	0.078
With context z	**0.041**	**0.070**

Table 1. experiment results

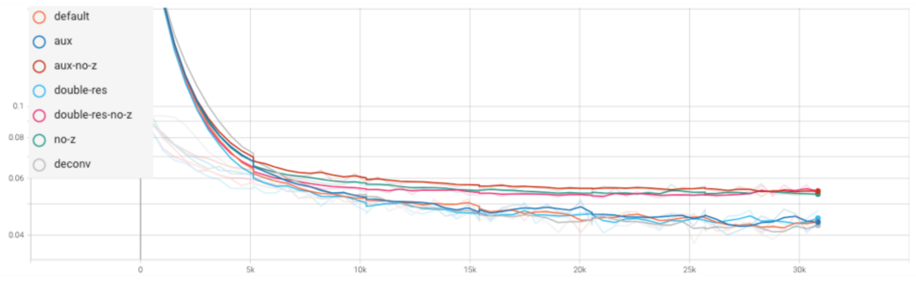

Fig. 5 Evaluation CE loss

In addition to the proposed architecture (fig. 4), we also tested other architectures including:

- MLP decoder (default in fig. 5)
- Using image segmentation as auxiliary loss (aux)
- Double resolution (double res)

The evaluation CE loss curve (Fig. 5) shows that architectures using contextual information consistently outperform those not using the contextual information.

A qualitative comparison between the proposed method, the baseline that without using contextual information, and the classic pipeline can be found in fig 6. Notice that in our proposed method, the smaller object (e.g. book and lamp) can be labelled correctly, whereas other methods failed to detect such objects. In addition, our method produces more accurate predictions for other objects (e.g. bed) because it can learn the regularities in object shapes.

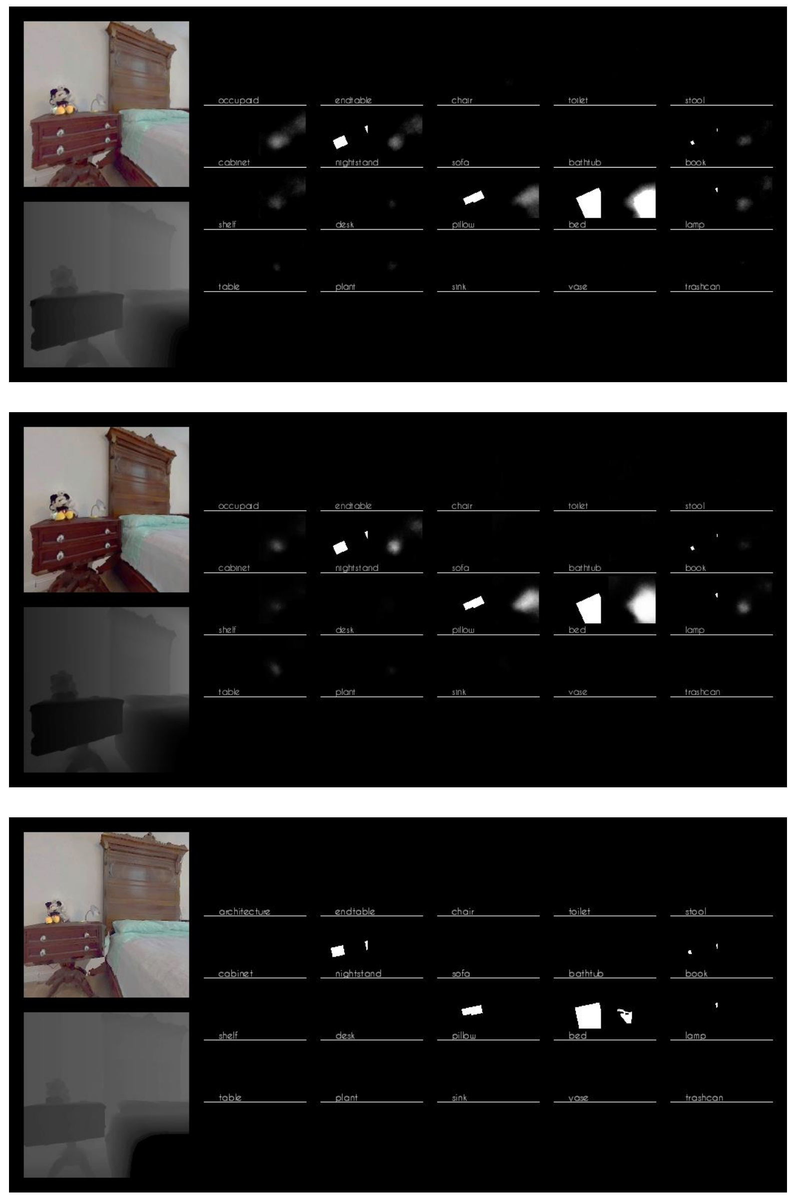

Fig. 6 Top: proposed method; middle: baseline w/o contextual information; bottom: classic pipeline. In all images, the left half of the object label is the ground-truth and the right half is the prediction by the model.