Semantic MapNet: Building Allocentric Semantic Maps and Representations from Egocentric Views

Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, Dhruv Batra

1 Georgia Institute of Technology
2 Oregon State University
3 Facebook AI Research
4 Google Research

{vcartillier3, zren80, tehaj, dbatra}@gatech.edu, leestef@oregonstate.edu

Abstract

We study the task of semantic mapping – specifically, an embodied agent (a robot or an egocentric AI assistant) is given a tour of a new environment and asked to build an allocentric top-down semantic map (‘what is where?’) from egocentric observations of an RGB-D camera with known pose (via localization sensors). Towards this goal, we present Semantic MapNet (SMNet), which consists of: (1) an Egocentric Visual Encoder that encodes each egocentric RGB-D frame, (2) a Feature Projector that projects egocentric features to appropriate locations on a floor-plan, (3) a Spatial Memory Tensor of size floor-plan length × width × feature-dims that learns to accumulate projected egocentric features, and (4) a Map Decoder that uses the memory tensor to produce semantic top-down maps. SMNet combines the strengths of (known) projective camera geometry and neural representation learning. On the task of semantic mapping in the Matterport3D dataset, SMNet significantly outperforms competitive baselines by 4.01 – 16.81% (absolute) on mean-IoU and 3.81 – 19.69% (absolute) on Boundary-F1 metrics. Moreover, we show how to use the neural episodic memories and spatio-semantic allocentric representations built by SMNet for subsequent tasks in the same space – navigating to objects seen during the tour (‘Find chair’) or answering questions about the space (‘How many chairs did you see in the house?’). Project page: https://vincentcartillier.github.io/smnet.html.

Introduction

Imagine yourself receiving a tour of a new environment. Maybe you visit a friend’s new house and they show you around (‘This is our living room, and down here is the study’). Or maybe you accompany a real-estate agent as they show you a new office space (‘These are the cubicles, and down here is the conference room’). Or someone gives you a tour of a mall or a commercial complex. In all these situations, humans have the ability to form episodic memories and spatio-semantic representations of these spaces (O’keefe and Nadel 1978). We can recall which spaces we visited (living room, kitchen, bedroom, etc.), what objects were present (chairs, tables, whiteboards, etc.), and what their relative arrangements were (the kitchen was combined with the open plan living room, the bedroom was down the hallway, etc.). We can also leverage these representations to perform new tasks in these spaces (e.g. navigate to the restroom via a path shorter than the one demonstrated on the tour). Of course, human memory is limited in time and in the level of metric detail it stores (Epstein et al. 2017). Our long-term goal is to develop super-human AI agents that can build rich, accurate, and reusable spatio-semantic representations from egocentric observations. This capability is an essential building block for autonomous navigation, mobile manipulation, and egocentric personal AI assistants.

In this paper, we study the specific task of creating an allocentric top-down semantic map of an indoor space, illustrated in Fig. 1. An embodied agent (a virtual robot or an egocentric AI assistant) is equipped with an RGB-D camera with known pose (extracted via localization sensors such as GPS and IMU). The agent is provided a tour of a new environment, represented as a trajectory of camera poses (shown in Fig. 1(a)). The task then is to produce an allocentric top-down semantic map (shown in Fig. 1(d)) from the sequence of egocentric observations with known pose (shown in Fig. 1(b)). Our experiments focus on top-down semantic segmentation, i.e. each pixel in the top-down map is assigned to a single class label (of the tallest object at that location on the floor, i.e. the one visible from the top-down view). Our produced semantic maps are metric – each pixel corresponds to a 2cm × 2cm grid on the floor – as opposed to topological maps (Fraundorfer, Engels, and Nister 2007; Nagarajan et al. 2020) that lack spatial information (scale, position) and do not support our downstream tasks of interest. Importantly, while the semantic top-down map is our...
primary ‘product’, our goal is to build neural episodic memories and spatio-semantic representations of 3D spaces in the process. Representations that enable the agent to easily learn and accomplish subsequent tasks in the same space – navigating to objects seen during the tour (‘Go to a chair’) or answering questions about the space (‘How many chairs did you see in the house?’).

What should we project? Approaches to top-down or overhead semantic segmentation can be arranged on spectrum illustrated in Fig. 2. At one end (on the right), are approaches (Sengupta et al. 2012; Sünderhauf et al. 2016; Maturana et al. 2018a) that first perform egocentric semantic segmentation and then use the known camera pose and the depth of each pixel to project labels to an allocentric map. In our experiments, we find that this results in ‘label splatter’ – any mistakes in the egocentric semantic segmentation made at the depth boundaries of objects get splattered on the map around the object. This problem can be slightly assuaged via image processing heuristics (median filtering). However, the fundamental issue persists even after those ‘bells and whistles’. Quantitatively, this results in high precision but low recall of the object segmentation. At the other end of the spectrum (on left) are approaches that operate on a single overhead image (projecting pixels if needed) and perform semantic segmentation on this image (Singh et al. 2018; Mátyus et al. 2015). While this may be appropriate for aerial or geospatial imagery, converting multiple high-res egocentric images into a single bird’s eye view is wasteful and throws out significant visual information. Qualitatively, we find that this results in coarse segmentations; object sizes are under-estimated, small objects missed entirely. Quantitatively, we see low precision and low recall.

We pursue an approach called Semantic MapNet (SMNet) that lies in the middle of this spectrum. Specifically, as shown in Fig. 2 (middle), SMNet extracts visual features in the egocentric reference frame, but predicts semantic segmentation labels in the allocentric reference frame. This is accomplished by projecting egocentric features to appropriate locations in an allocentric spatial memory, and using this memory to decode a top-down semantic segmentation. This design addresses both deficiencies in prior work – (a) the spatial-memory-to-map decoder in SMNet is based on transposed convolutions and learns to smooth out any ‘feature splatter’; and (b) the egocentric feature extractor in SMNet operates directly on high-res egocentric images and is able to recognize and segment small objects that may not be visible from a bird’s eye view.

We conduct experiments on the photo-realistic scans of building-scale environments (homes, offices, churches) in the Matterport3D dataset (Chang et al. 2017) using the Habitat simulation platform (Savva et al. 2019) (giving us access to agent state, navigation trajectories, RGB-D renderings, etc.). We choose the Matterport3D dataset because it provides semantic annotations in 3D, the spaces are large enough to allow multi-room traversal by the agent, and the use of a 3D simulator allows us to render RGB-D from any viewpoint, create top-down semantic annotations, and study embodied AI applications in the same environments. Quantitatively, on the task of semantic mapping, SMNet significantly outperforms the aforementioned baselines by 4.01 – 16.81% (absolute) on mean-IoU and 3.81 – 19.69% (absolute) on Boundary-F1 metrics.

SMNet combines the strengths of (known) projective camera geometry with neural representation learning, and address our key desideratum – learning rich, reusable spatio-semantic representations. We demonstrate via extension experiments how representations built by SMNet from a single tour of an environment can be reused for ObjectNav and Embodied Question Answering (Das et al. 2018).

Related Work

Spatial Episodic Memories for Embodied Agents. Building and dynamically updating a spatial memory is a powerful inductive bias that has been studied in many embodied settings. Most SLAM systems perform localization by registration to sets of localized keypoint features (Mur-Artal and Tardós 2017). Many recent works in embodied AI have developed agents for navigation (Anderson et al. 2019; Beeching et al. 2020; Gupta et al. 2017; Georgakis, Li, and Kosecka 2019; Blukis et al. 2018) and localization (Henriques and Vedaldi 2018; Parisotto and Salakhutdinov 2017; Zhang et al. 2017) that build 2.5D spatial memories containing deep features from egocentric observation. Like our approach, these all involve some variation of egocentric feature extraction, pin-hole camera projection, and map update mechanisms. However, these works focus on spatial memories as part of an end-to-end agent for a downstream task and do not evaluate the quality of the generated maps in terms of environment semantics directly, nor study how segmentation quality affects downstream tasks.

Semantic Mapping from Egocentric Observations. Predicting top-down semantic segmentation from egocentric observations has been studied in the context of robotics as the semantic SLAM (or semantic mapping) problem (Risnol et al. 2019; Maturana et al. 2018b; Grinvald et al. 2019; McCormac et al. 2017). We compare with a recent representative algorithm in this family as our baseline (Grinvald et al. 2019). Further work has examined the use of semantic labels as an intermediate step in an end-to-end model (Gordon et al. 2018; Chaplot et al. 2020a) or to derive supervision to reward agent trajectories (Chaplot et al. 2020c). These works have not evaluated the quality of the semantic map and instead focused on downstream tasks. All these works follow the Segment-then-Project paradigm – invoking a segmenta-
Semantic MapNet (SMNet)

We now describe our proposed approach for semantic mapping, called Semantic MapNet (SMNet), in detail. As shown in Fig. 3, SMNet consists of the following modules:

- An **Egocentric Visual Encoder** that converts each egocentric RGB-D frame into a $R^{w \times h \times d}$ feature tensor, representing the content of each image region.
- A **Feature Projector** that uses the known camera pose and the depth of each pixel to project these egocentric features to appropriate locations on a floor-plan.
- A **Spatial Memory Tensor** of size floor-plan length $\times$ width $\times$ feature-dims that accumulates these projected egocentric features. Repeated observations of the same spatial locations are incorporated through a learned recurrent model operating at each location.
- A **Map Decoder** that uses the accumulated memory tensor to produce top-down semantic segmentations.

**Problem Setup and Notation.** Let $I$ denote an RGB-D image. We assume a known camera – specifically, let $K$ be the camera intrinsic matrix, and $[R | t]$ denote the camera extrinsic matrix (rotation and translation needed to convert world coordinates to camera coordinates). Thus, an agent’s trajectory through an environment is represented as a sequence of egocentric RGB-D observations $I^{(1)}, \ldots, I^{(T)}$ at known poses $[R | t]^{(1)}, \ldots, [R | t]^{(T)}$. Strictly speaking, our approach does not require knowing camera pose in world coordinates at all times – all we need are successive pose transformations $[R | t]^{(t \rightarrow t+1)}$, a problem known in robotics and computer vision as egomotion estimation. The entire approach could be defined in terms of the camera coordinates at time $t = 1$. However, for sake of clarity of the exposition, we describe our approach with global pose.

Let $S$ denote the top-down semantic segmentation. Each pixel in $S$ represents a $2cm \times 2cm$ cell in the environment and is labeled with the class of the tallest object in that cell (i.e. the object visible from above). At each time $t$, let $M^{(t)}$...
denote the memory tensor incorporating all the information observed in the trajectory so far, and let \( S^{(t)} \) denote the segmentation predicted using \( M^{(t)} \). Note that test-time evaluation is done using \( S^{(T)} \), but during training our agent predicts and receives supervision for intermediate predictions along the trajectory \( S^{(t)} \), \( t = 1, \ldots, T \).

There are a number of coordinate systems in this discussion which we define now for clarity – pixel positions in the egocentric RGB-D image \( I \) are indexed with \( i, j \), and the depth at this pixel is denoted with \( d_{i,j} \) (or \( d \) when its clear from context which pixel is being talked about). A 3D point in world coordinates is denoted with \( x, y, z \). For notational simplicity, we follow the standard convention in computer graphics – negative Y-axis aligned gravity in the world coordinate system. Finally, cells in the memory tensor are indexed with \( u, v \). Next, we describe each module in detail.

**Egocentric Visual Encoder.** Each egocentric frame \( I^{(t)} \) gives a local glimpse of the environment – providing information about objects and their locations in the current view. To represent these, we encode each RGB-D frame using RedNet (Jiang et al. 2018), a recently proposed architecture for semantic segmentation of indoor scenes. In principle, one may choose any standard image encoder network for semantic segmentation such as Mask-RCNN (He et al. 2017). We chose RedNet simply because the network structure has proven to be effective for parsing indoor environments and pre-trained models (learned on SUN-RGBD dataset (Song, Lichtenberg, and Xiao 2015)) are publicly available. We initialize with these pre-trained weights and fine-tune RedNet on our dataset. We conducted several experiments by extracting egocentric features at different stages in the RedNet network (encoder, last layer, scores, softmax, one-hot encoded labels). We found that encoding each RGB-D frame with the last layer RedNet features yields to the best performances. The output of this encoder for image \( I^{(t)} \) is an egocentric feature map \( F^{(t)} \) \( \in \mathbb{R}^{240 \times 320 \times 64} \) with each of the 240 \( \times \) 320 cells storing a 64-d feature. We upscale this tensor to the resolution of the depth image \( (480 \times 640) \) with bilinear interpolation, resulting in each pixel \( i, j \) having an associated feature \( F_{i,j}^{(t)} \) \( \in \mathbb{R}^{64} \) and depth value \( d_{i,j} \).

**Feature Projector.** To project an egocentric feature \( F_{i,j}^{(t)} \) to the spatial memory, we must (a) shoot a ray from the camera center through the image pixel \( (i, j) \) out to a depth \( d_{i,j} \) to get a 3D point in the camera coordinate system, (b) convert from camera to world coordinates to get the corresponding \( (x, y, z) \), and then (c) project it to cell indices \( u, v \) in the memory tensor. With known camera pose and intrinsics, these transformations for the standard pinhole camera can be written compactly as:

\[
\begin{bmatrix}
x \\
y \\
z \\
\end{bmatrix} = d_{i,j} R^{-1} K^{-1} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} - t, \quad \text{and} \quad \begin{bmatrix} u \\ v \\ 0 \\ 1 \end{bmatrix} = P_v \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}
\]

where \( P_v \) is a known orthographic projection matrix converting 3D world coordinates to 2D memory cell indices. When several points are projected to the same index in \( M^{(t)} \), we retain the one with the maximum height in the world coordinates. This results in a set of projected features \( F_{u,v}^{(t)} \).

**Spatial Memory Tensor** \( M \) is a 3D tensor of size \( U \times V \times 256 \). Each grid cell \( (u, v) \) stores a 256-d feature vector and corresponds to a 2cm \( \times \) 2cm area on the floor-plan, which is the same spatial resolution as the segmentation \( S \). The memory must be updated at each time step to incorporate new observations. Specifically, given the current memory \( M^{(t-1)} \) and a new observation \( F_{u,v}^{(t)} \) for cell \( u, v \), we compute \( M_{u,v}^{(t)} = GRU(F_{u,v}^{(t)}, M_{u,v}^{(t-1)}) \) where \( M_{u,v}^{(t-1)} \) is the hidden state and \( F_{u,v}^{(t)} \) the input for the GRU. This GRU can learn to accumulate incoming observations. Notice that the GRU parameters are shared spatially, i.e. for all \( (u, v) \). Importantly, this independent updating of modified memory cells (as opposed to something like a ConvGRU) ensures that observations only affect local regions of the memory – keeping previously observed areas stable.

**Map Decoder.** The memory tensor \( M^{(t)} \) is used to decode a top-down semantic segmentation map. We use a simple architecture consisting of five convolutional layers with batch norm and ReLU activations. As the memory \( M \) and segmentation \( S \) are the same spatial resolution, no learned upsampling or downsampling is involved in this decoding.

Together, these modules form SMNet and implement the basic principle of ’encode pixels, project features to a spatial memory, decode labels’. Notice that all modules and thus the entire architecture is end-to-end differentiable.

**Matterport Semantic-Map Dataset**

For our experimental evaluation, we need 3D environments for an agent to traverse that have dense semantic segmentations. While our extension experiments involve agent-driven navigation, our core task of semantic mapping is defined w.r.t. a fixed trajectory provided to the agent as input. The more challenging task of simultaneous semantic mapping and goal-driven navigation is left for future work.

Given this fixed-trajectory setting, our task has the input (but not output) specification of video segmentation – both taking a sequence of input images along a trajectory. We choose the Matterport3D scans (Chang et al. 2017) with the Habitat simulator (Savva et al. 2019) over video segmentation datasets for a number of reasons – Matterport3D provides semantic annotations in 3D (as opposed to 2D annotations in video datasets), the spaces are large enough to allow multi-room traversal by the agent (as opposed to (Dai et al. 2017; Nathan Silberman and Fergus 2012), and the use of a 3D simulator (as opposed to (Geiger, Lenz, and Urtasun 2012; Cords et al. 2016)) allows us render RGB-D from any viewpoint, create top-down semantic annotations, and study embodied AI applications in the same environments.

**Matterport3D Environments.** Matterport3D dataset (Chang et al. 2017) contains reconstructed 3D meshes of 90 indoor environments (homes, offices, churches). These meshes are densely annotated with 40 object categories. Many of these are rare or would appear
as thin lines in a top-down view (e.g., walls and curtains); we focus on the 12 most common object categories: chair, table, cushion, cabinet, shelving, sink, dresser, plant, bed, sofa, counter, fireplace (sorted in descending order by number of object instances). We treat all other classes and background pixels as void class. We divide multi-story environments in Matterport3D into separate floors by manually refining floor dividers present in the meta-data. This is not always possible given a single dividing plane (e.g., split level homes), resulting in inaccurate top-down maps—we discard such environments. Utilizing the same data split as (Wijmans et al. 2019), we keep 85 unique floors in our dataset: 61 for training, 7 for validation, and 17 for testing. See supplement (Cartillier et al. 2020) for these splits.

**Ground-Truth Top-down Semantic Segmentations.** To supervise our model, we need access to ground-truth top-down semantic maps from these environments. These are created by applying an orthographic projection for the 3D mesh annotations in a similar manner to Eqn. 1 (right). In this process we only project vertices labeled with one of the 12 kept object categories. The resulting ground-truth top-down semantic maps are free from occlusions caused by non-target objects. There will be cases where from the egocentric view the agent won’t be able to visualize the object entirely either because it is occluded or the object is too high (wall cabinet). In table 1 we report numbers on the Seg. GT → Proj. experiment where the agent projects egocentric ground-truth semantic labels. This will account for such occlusion and set an upper-bound to our experiments.

**Modal Maps and Viewing Frustum.** We perform modal top-down semantic segmentation (as opposed to amodal). Specifically, the agent receives supervision on map cells it has actually observed: it is not evaluated on hallucinating unseen regions. We do this by projecting the viewing frustum (i.e., region the agent can currently see) to the floor-plan at each navigation step. We can then keep track of which regions have been observed during a trajectory.

**Navigation Paths.** We assume that an agent’s path through the environment is provided by some external policy—e.g., a goal-oriented path or a general exploration policy—and that we are constructing the map and memory opportunistically from this experience. To simulate this behavior, we manually record a navigation path through each floor using the Habitat simulator (Savva et al. 2019). The action space is move forward 10cm, and rotate left or right $9^\circ$. To encourage trajectories with high environment coverage, our human navigation interface included the top-down RGB map with agent position drawn. On average, agents move 2500 steps in each environment. Note that this is an order of magnitude longer than most navigation trajectories in contemporary works (Savva et al. 2019; Wijmans et al. 2020; Kadian et al. 2019; Gordon et al. 2018; Chaplot et al. 2020b).

**Training Samples.** To train our model, we consider 20-step navigation segments from these trajectories. Starting from a random location on the trajectory, we step the agent forward 20 steps along it, capturing the corresponding viewpoints to mask the top-down semantic map. We generate 50 examples for each environment leading to 3050/350 training samples. This both greatly increases the number of training instances and increases training speed by requiring a smaller semantic memory tensor. At evaluation/testing time, the agent builds the map from the entire trajectory.

**Semantic Mapping Experiments**

**Baselines.** As depicted in Fig. 2, there exists a spectrum of methodologies for our task based on what is being projected from egocentric observations to the top-down map—pixels, features, or labels. Our approach stakes a middle-ground on
Implementation Details

We pretrain two RedNet (Jiang et al. 2018) models for semantic segmentation in our setting – one from egocentric RGB-D (Segment → Project) and another from top-down RGB alone (Project → Segment). SMNet. We use a single-layer GRU to update the spatial memory. SMNet is trained end-to-end under cross-entropy loss using SGD with learning rate $10^{-4}$, momentum 0.9, weight decay $4e^{-4}$, and batch size 8 across 8 Titan XPs. Training took 2-3 days. Back propagation is applied after 20 steps.

Evaluation Metrics. We report the entire range of evaluation metrics for semantic segmentation: (a) the overall pixelwise labeling accuracy (Acc), (b) the average of pixel recall or precision scores for each class (mRecall/mPrecision), (c) the average of the intersection-over-union score of all object categories (mIoU), and (d) the average of the boundary F1 score of all object categories (mBF1). mBF1 is contour-based metric in (Csurka et al. 2013). mIoU and mBF1 serve as our primary metrics.

Results. Table 1(left) shows a summary of the results with bootstrapped standard error (see supplement (Cartillier et al. 2020) for category-level breakdowns). Fig. 4 shows qualitative results. Project → Segment achieves low performance (mBF1 17.33, mIoU 19.96) compared to the approaches that operate on egocentric images prior to projection (either via segmentation or feature extraction). This suggests details lost in the top-down view are important for disambiguating objects – e.g., the chairs at the table in Fig. 4 (bottom) are difficult to see in the top-down RGB and are completely lost by this approach. Segment → Project performs significantly better (mBF1 33.21, mIoU 32.76), but faces a problem with errors in the egocentric predictions resulting in noise in the top-down map. Semantic SLAM (VoxelBlox++) performs worse than the Segment → Project baseline (mBF1 31.05, mIoU 28.11). VoxelBlox++ follows a segment-then-project paradigm, making it prone to the same errors as the Segment → Project baseline. In addition, the data association module of VoxelBlox++ will sometimes group objects of different categories (e.g., the two bottom chairs are grouped with the table in Fig. 4). As our approach reasons over a spatial memory tensor, it can reason about multiple observations of the same point – achieving mBF1 37.02, and mIoU 36.77. The Segment GT → Project experiment sets an upper-bound of performances (perfect pre-
Figure 5: Object Navigation: Visualization of paths found by A* using SMNet maps. Green and red squares indicate agent’s starting and stopping locations. The grey color represents the floor pixels. The left example shows a case of success with high SPL = 0 and the example on the right shows a case of success with low SPL = 0.4282.

Figure 6: Visualizations of MemoryQA

a rich description of the space. In this section, we explore proof-of-concepts for various downstream embodied AI tasks based on these representations.

**Object Navigation.** A natural extension is navigating to specific objects, or ObjectNav for short. In ObjectNav, an agent is randomly initialized in a scene and tasked to navigate to an instance of a given object class as quickly as possible (Savva et al. 2019). In the standard setting, the environment is novel; however, we consider a pre-exploration setting where the agent first traverses the environment to construct a top-down semantic map. In parallel we compute a top-down map of heights and use it to compute a free space map of the environment. We opt for an open loop planning strategy by running A* search (with a Euclidean heuristic) on the free space map combined with the semantic map to find a path from the start location to the nearest target object instance and then run the trajectory in the Habitat simulator (Savva et al. 2019). We evaluated this strategy on the validation set of the ObjectNav Habitat challenge (hab 2020), suggesting that the memory tensor contains useful spatial and semantic information in this pre-exploration setting. Experimentally we found that inaccuracies in the free space map computation and objects misclassification in the top-down semantic map are the two major sources of error. While the latter is harder to cope with, the former can be limited by extended SMNet to predict free space. Fig. 5 shows qualitative results of two successful examples – start locations are shown as green squares with trajectory transitioning to red until terminating. Using the predicted semantic maps provides interpretability – when the navigation fails, we can know why. On the example on the right in Fig. 5 the chair at the top has been mislabeled as sofa, thus leading the agent to the second closest chair slightly on the left.

**Question Answering.** We also consider an embodied question answering (Das et al. 2018) task where agents are asked questions about the environment. Again considering a pre-exploration setting, the agent first navigates the environment on a fixed trajectory to generate the spatial memory tensor. We consider counting questions (e.g., ‘How many beds are there?’) and design a decoder directly from the spatial memory. The decoder outputs the number of instances detected per object category for a given memory input. We train this decoder using 5m x 5m memory samples. We design this task as a classification problem with 21 classes corresponding to values ranging from 0 to 19 and 20+. When testing on larger environments, we apply this decoder using a sliding window over the full memory – accumulating counts. We compare our approach to a ‘prior’ baseline that answers with the most frequent answer in the training set. Our approach outperforms this baseline across the board: 27.78% vs. 20.83% on accuracy, 13.19% vs. 9.18% class-balanced accuracy and 5.35 vs. 6.98 on RMSE.

**Conclusion**

Taken holistically, our results show SMNet is able to outperform competitive baselines in constructing semantic maps, and spatio-semantic representations built show promise on downstream tasks. Note that the specific sub-task of counting instances highlights a limitation in our current problem setup – using semantic segmentation does not preserve instance information. The generalization to producing top-down instance segmentation maps is an interesting avenue for future work.
References


