KNOWLEDGE-DRIVEN SCENE PRIORS FOR SEMANTIC AUDIO-VISUAL EMBODIED NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalisation to unseen contexts remains a challenge for embodied navigation agents. In the context of semantic audio-visual navigation (SAVi) tasks, generalisation includes *both* generalising to unseen indoor visual scenes as well as generalising to unheard sounding objects. Previous SAVi task definitions do not include evaluation conditions on truly novel sounding objects, resorting instead to evaluating agents on unheard sound clips of known objects; meanwhile, previous SAVi methods do not include explicit mechanisms for incorporating domain knowledge about object and region semantics. These weaknesses limit the development and assessment of models' abilities to generalise their learned experience. In this work, we introduce the use of knowledge-driven scene priors in the semantic audio-visual embodied navigation task: we combine semantic information from our novel knowledge graph that encodes object-region relations, spatial knowledge from dual Graph Convolutional Networks, and background knowledge from a series of pre-training tasks—all within a reinforcement learning framework for audio-visual navigation. We define a new audio-visual navigation sub-task, where agents are evaluated on novel sounding objects, as opposed to unheard clips of known objects. We show state-of-the-art results on multiple semantic audio-visual navigation benchmarks, within the Habitat-Matterport3D simulator, where we also show improvements in generalisation to unseen regions and novel sounding objects. We release our code, knowledge graph, and dataset in the supplementary material.

1 INTRODUCTION

Humans are able to make use of background experience, when navigating unseen or partially-observable environments. Prior experience informs their world model of the semantic relationships between objects commonly found in an indoor environment, the likely object placements, and the properties of the sounds those objects emit throughout their object-object and object-scene interactions. Artificial embodied agents, constructed to perform goal-directed behaviour in indoor scenes, should be endowed with similar capabilities; indeed, as autonomous agents enter our homes, they will need intuitive understanding about how objects are placed in different regions of houses, for better interaction with the environment. Whereas external (domain) knowledge can yield improvements in agent sample-efficiency while learning, generalisability to unseen environments during inference, and overall interpretability in its decision-making, the goal of finding generalisable solutions for injecting knowledge in embodied agents remains elusive (Oltramari et al., 2020; Francis et al., 2021).

The task of audio-visual navigation lends itself especially well to the use of domain knowledge, e.g., in the form of human-inspired background experience (encapsulated as a prior over regions and semantically-related objects contained therein). Certain sounds can be associated with particular places, e.g., a smoke alarm is more likely to originate in the kitchen, while telephone ringing sound is more likely to come from the office. To infer such semantic information from sound inputs in an environment, we propose the idea of knowledge-enhanced prior.

By using a prior enriched with general experiences, we hypothesise that the learned model would generalize to novel sound sources. We adopt a modular training paradigm, which has been shown to lead to improvements in cross-domain generalizability and more tractable optimisation (Chen et al., 2021b; Chaplot et al., 2020b; Francis et al., 2021). To verify our hypothesis on generalizability, we

evaluate the agent's performance on a set of novel sounding objects that were not introduced during training.

Contributions. First, we introduce the use of knowledge-driven scene priors in the semantic audio-visual embodied navigation task: we combine semantic information from our novel knowledge graph that encodes object-region relations, spatial knowledge from dual Graph Convolutional Networks, and background knowledge from a series of pre-training tasks—all within a reinforcement learning framework for audio-visual navigation. Second, we define a knowledge graph that encodes object-object, object-region, and region-region interactions in a photorealistic 3D indoor navigation environment; we use this knowledge graph to pre-train specialised components of our modular framework, for goal specification and progress-monitoring. Next, we curate a multimodal dataset for additional pre-training of a visual encoder, in order to encourage object-awareness in visual scene understanding. Finally, we define a new task of semantic audio-visual navigation, where we assess agent performance on the basis of their generalisation to novel sounding objects. We offer experimental results, against strong baselines, and show improvements over these models on various performance metrics in unseen contexts. We provide all code, dataset-generation utilities, and knowledge graph in the supplementary.

2 Related Work

Modularity in goal-driven robot navigation. Goal-oriented navigation tasks have long been a topic of research in robotics (Kavraki et al., 1996; Lavalle et al., 2000; Canny, 1988; Koenig & Likhachev, 2006). Classical approaches generally tackle such tasks through non-learning techniques for searching and planning, e.g., heuristic-based search (Koenig & Likhachev, 2006) and probabilistic planning (Kavraki et al., 1996). Although classical approaches might offer better generalisation and optimality guarantees in low-dimensional settings, they often assume accurate state estimation and cannot operate on high dimensional raw sensor inputs (Gordon et al., 2019). More recently, researchers have geared toward data-driven techniques, e.g., deep reinforcement learning (Wijmans et al., 2020a; Batra et al., 2020; Chaplot et al., 2020a; Yang et al., 2019; Chen et al., 2021b;a; Gan et al., 2020) and imitation learning (Irshad et al., 2021; Krantz et al., 2020), to design goal-driven navigation policies. End-to-end mechanisms have proven to be powerful tools for extracting meaningful features from raw sensor data, and thus, are often favoured for this type of setting where agents are tasked with learning to navigate toward goals in unknown environments using mainly raw sensory inputs. However, as task complexity increases, this type of systems generally exhibit significant performance drops specially in unseen scenarios and in long-horizon tasks (Gordon et al., 2019; Saha et al., 2021).

To address the aforementioned limitations, modular decomposition has been explored in recent embodied tasks. Chaplot et al. (2020c) design a modular approach for visual navigation consisting of a module that builds and updates a map of the environment, and a global and local policies to, respectively, predict the next sub-goal using such map and the low-level actions to reach it. Irshad et al. (2021) also leverage a hierarchical setup to disentangle Vision-Language Navigation (VLN) (Anderson et al., 2018b) into a global policy tasked with grounding the input modalities and predicting the next global step, and a local policy that performs motion control to navigate toward it. Gordon et al. (2019) design a hierarchical controller that invokes different low-level controllers in charge of different tasks such as planning, exploration and perception. Similarly, Saha et al. (2021) design a modular mechanism for Vision-Language tasks that breaks down the task into multiple sub-tasks that include: mapping, language understanding, modality grounding, and planning. The aforementioned modular designs have shown to increase task performance and generalisability, especially in unexplored scenarios, compared to their end-to-end counterparts. Motivated by the aforementioned, we develop a modular framework for semantic audio-visual navigation, which includes pre-trained and knowledge-enhanced scene priors, which enable improved unseen generalisation.

Knowledge graphs in visual navigation. Combining prior knowledge with machine learning systems remains a widely-investigated topic in various research fields, such as natural language processing (Ma et al., 2021; 2019; Francis et al., 2021), due to the improvements in generalisability and sample-efficiency that symbolic representation promises for learning-based approaches. Historically, integrating symbolic knowledge with, e.g., navigation agents has proven non-trivial, yielding a collection of research areas focusing on smaller components of the problem—such as finding the appropriate representation of the knowledge (e.g., logical formalism, knowledge graphs, proba-

bilistic graphical models), the appropriate *type* of knowledge that should be encoded (e.g., spatial commonsense, declarative facts, etc.), and the best knowledge injection mechanism (e.g., graph convolutional networks, grounded natural language, etc.) (Ma et al., 2019). Knowledge graphs have gained popularity, due to their interpretability and general availability as existing large-scale resources, such as ConceptNet (Speer et al., 2016) and VisualGenome (Krishna et al., 2016a). Fortuitously, graph processing of structured data has experienced a surge of popularity in deep learning, in recent years, leading to renewed interest in this neuro-symbolism (Oltramari et al., 2020; Wu et al., 2021). Some works in visual navigation tasks exploit knowledge graphs, in the pursuit of generalisation (Moghaddam et al., 2020; Yang et al., 2019; Lv et al., 2020; Du et al., 2020; Vijay et al., 2019). Yang et al. (2019) create knowledge graphs based on the VisualGenome (Krishna et al., 2016b) and inject features extracted from the graph as prior knowledge in visual navigation. In similar fashion, Qiu et al. (2020) provide agents with knowledge of object-object relational semantics. Lv et al. (2020) show improvements in goal-directed visual navigation, by injection 3D spatial knowledge into learning-based agents. Inspired by these works, we construct a knowledge graph that includes *both* object-object and object-region semantics, which enables the more complex reasoning path, *sound* $\rightarrow$ *object* $\rightarrow$ *region*, in the audio visual navigation task. To our best knowledge, we therefore become the first to study knowledge-driven scene priors for the audio-visual navigation task family.

Generalization to unseen contexts. Chen et al. (2020; 2021b;a) leverage the SoundSpaces (Chen et al., 2020) simulation environment and dataset to design and assess Audio-Visual Navigation policies. The dataset is based on photorealistic indoor environments from the Matterport3D (Chang et al., 2017) and Replica (Straub et al., 2019) datasets, to which 102 sound sources commonly found in indoor environments (e.g., household appliances, musical instruments, telephones, etc.) were incorporated. The SoundSpaces dataset is split, such that indoor scenes encountered during testing are not found in the episodes used during the training stage. However, sounds of objects encountered during training may also appear during testing. Gan et al. (2020) also explore Audio-Visual Navigation, but using the simulation platform AI2-THOR (Kolve et al., 2017). The authors introduce the Visual-Audio Room (VAR) benchmark consisting of seven different indoor environments—two of which were used for training and five for testing. The VAR benchmark incorporates three different audio categories: ring tone, alert alarm, and clocks. Similar to the AVN task introduced before, sound sources are found both in the training scenes, as well as and the testing scenes. In this paper, we argue that in the context of Audio-Visual Navigation tasks, generalisation to unseen environments pertains to both generalising to unseen visual scenes, as well as to unheard sounds. Current Audio-Visual benchmarks do not take into consideration the latter. Thus, there is no direct assessment of generalisation performance to unheard sounds. To tackle this limitation, we propose a curated version of the SoundSpaces dataset where we evaluate our agent in two different settings: (1) seen scenes and unheard sounds, (2) unseen scenes and unheard sounds.

3 Problem Definition

We consider the semantic audio-visual navigation (S-AVN) task proposed by Chen et al. (2021a). In this task, the agent is initialised at a random location, in an unmapped 3D house environment, containing a sounding object (e.g., piano). The agent's task is to reach the sounding object using its sensory inputs, consisting of visual and audio sensors. Two assumptions are made in this task: 1) the target sound has a variable length and may not be available at each time step, so the sound may stop during navigation (e.g., telephone ringing sound stops after some time); 2) the sounding object has a visual embodiment, which is semantically meaningful (e.g., the sound produced by a spoon dropping is associated with the spoon). These assumptions are realistic because sound events have a variable length in the real world based on the semantics of the sounding object. For example, the sound produce by a glass jar breaking would usually be shorter than a telephone ringing sound. Due to the variable length nature of the sound, the agent cannot rely on the audio signal alone to reach the sounding object. Instead, the agent needs to use the audio signal to predict its location and understand the sounding object's semantics. Moreover, the agent also needs to use the visual cues for associating it with the sound semantics and reason about the object and region semantics to navigate effectively.

We further extend the S-AVN task by evaluating the agent on unheard (or novel) sounding objects. In the initial task (Chen et al., 2021a), the agent was evaluated on unheard clips of the known sounding objects, whereas in our task, the agent is evaluated on completely unknown sounding objects. More formally, let $\mathcal{H}$ be the set of houses, let $\mathcal{O}$ be the set of sounding objects (e.g., shower, tv monitor),

and let $\mathcal{R}$ be the set of regions (e.g., bathroom, living room). A house $h_i \in \mathcal{H}$ has a set of regions $(r_{i1}, r_{i2}, \ldots, r_{ij})$ and a set of objects $(o_{i1}, o_{i2}, \ldots, o_{ik})$, where there are k objects placed in j regions of the house h_i. Note that there are multiple instances of each sounding object $o \in \mathcal{O}$ and region $r \in \mathcal{R}$ across all houses $\mathcal{H}$. We divide the total set of possible houses $\mathcal{H}$ into two mutually exclusive subsets: $\mathcal{H}_{seen}$ and $\mathcal{H}_{unseen}$. Similarly, we divide sounding objects $\mathcal{O}$ into two subsets: $\mathcal{O}_{heard}$ and $\mathcal{O}_{unheard}$. The houses in $\mathcal{H}_{seen}$ and the sounding objects in $\mathcal{O}_{heard}$ are only experienced by the agent during the training phase; the agent is evaluated on unheard sounding objects $\mathcal{O}_{unheard}$. Thus, the agent must learn to reason about the novel sounds based on prior knowledge to solve this task. Our work aims to enable the agent to reach the sounding object it has never experienced before.

4 KNOWLEDGE-DRIVEN SCENE PRIORS FOR AUDIO-VISUAL NAVIGATION

4.1 SEMANTIC AUDIO-VISUAL EMBODIED NAVIGATION

We introduce a knowledge-driven approach for semantic audio-visual embodied navigation (`K-SAVEN`). `K-SAVEN` incorporates scene priors in knowledge graph form and extracts relational features using Graph Convolutional Network (GCN) (Kipf & Welling, 2017) for audio and visual modalities. GCN provides the agent reasoning capability using prior knowledge and dynamically updates its belief according to the current observation, specific to the current environment. Our model also incorporates Scene Memory Transformer (SMT) (Fang et al., 2019) that captures long-term dependencies by recording visual features in memory and locating the goal by attending to acoustic features. We use visual observations to compute visual features, including vision-based semantic knowledge vector and features encoded from the vision encoder. Similarly, we use audio observations to compute acoustic features, including audio-based semantic knowledge vector and location prediction from location predictor. Thus, the prior knowledge-driven reasoning capability using GCNs, with memory-based attention mechanism using SMT allow the agent to generalise to novel houses and sounding objects, exploit spatio-temporal dependencies, and navigate to the goal efficiently.

The `K-SAVEN` policy, as shown in Figure 1, consists of 5 modules: 1) Pre-trained models that, given the audio and visual observations from the environment, predict objects and regions; 2) Graph Convolutional Networks that compute audio-semantic and visual-semantic feature embeddings; 3) Vision Encoder that projects the visual observations at each step to an embedding space; 4) Location Predictor that, given the acoustic signal from the sounding object, predicts its relative distance and direction from the agent; 5) Scene Memory Transformer that uses an attention-based policy network, which computes a distribution over actions, given the encoded observations in scene memory and the current observation that captures goal information from acoustic events. In the following sections, we discuss each module in more detail.

4.1.1 MODULAR PRE-TRAINING

In our task, the agent relies on audio observations to set its goal and uses visual observation to navigate to that goal. Therefore, the agent must detect objects and regions in a given observation. To this end, we trained audio (f_c^a) and vision (f_c^v) classification models to predict classification scores for objects and regions in a given observation. More specifically, f_c^a and vision f_c^v predict a score for each object $o \in \mathcal{O}$ (the likelihood that the object o produced the observation) and region $r \in \mathcal{R}$ (the likelihood that the observation correspond to region r). These models are used as a backbone for the other models in our proposed architecture.

The acoustic event has variable length and may not be present at each time step, so the agent cannot rely on the current audio observation, alone, as a persistent signal. Thus, our model aggregates the current prediction $\hat{c}_t^a$ with the previous prediction c_{t-1}^a, $c_t^a = f_\lambda(\hat{c}_t^a, c_{t-1}^a) = (1-\lambda)\hat{c}_t^a + \lambda c_{t-1}^a$, where λ is the weighting factor set to 0.5. When the acoustic event stops (i.e., zero sound intensity), the agent uses its latest estimate c_t^a.

4.1.2 KNOWLEDGE GRAPH CONSTRUCTION

Our knowledge graph captures the relationships of object-to-object and object-to-region. It is denoted by undirected graph $G = (V, E)$, where V and E denote vertices and edges respectively. Here, each vertex denotes an object label or a type of room. We employ 21 object categories and 24 region

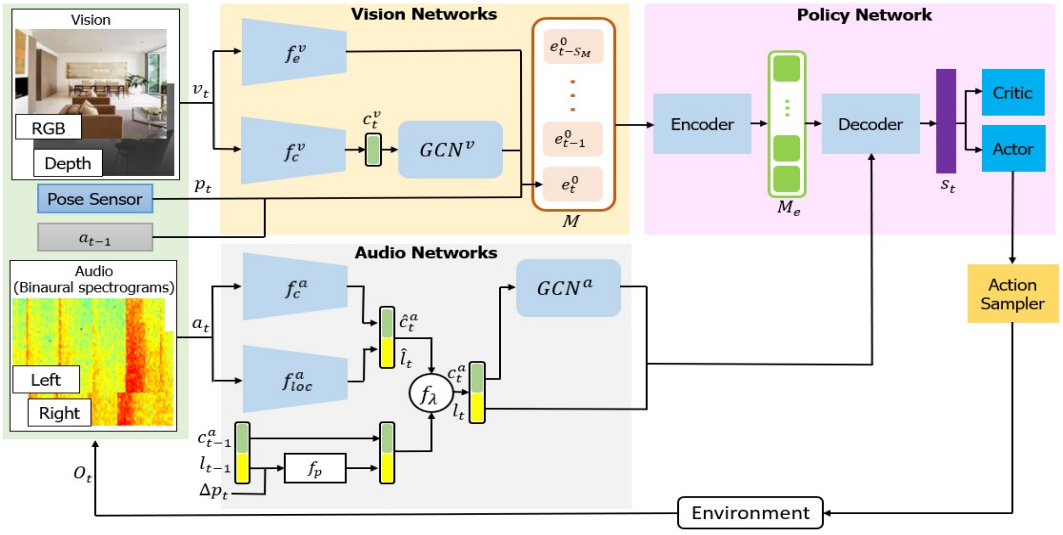

Figure 1: K-SAVEN's system overview. Visual observation v_t is fed to two modules: vision encoder f_e^v, which encodes the visual observation, and pre-trained vision model f_c^v, which, given the visual observation, predicts classification scores c_t^v for objects and regions. These scores are used by the vision-based graph convolutional network GCN^v to compute visual-semantic feature embeddings. The outputs of these two models are stored in memory M. Audio observation a_t is also fed to two models: location predictor f_{loc}^a, which predicts distance and direction of the sounding object from the agent (l_t), and pre-trained audio model f_c^a, which, given the audio observation, predicts classification scores c_t^a for objects and regions. These scores are used by the audio-based graph convolutional network GCN^a to compute audio-semantic feature embeddings. The attention-based policy network conditions the encoded visual information M_e on the acoustic information, enabling the agent to associate visual cues with acoustic events and predict the state representation s_t, which contains spatial and semantic cues helpful to reach the goal faster. The actor-critic network, given the state s_t, predicts the next action a_t. When the agent executes the action in the environment, it receives a reward and observations.

categories; its detail is described explicitly in the subsection 5.1. Edges represent occurrence of object-to-object or object-to-region; edges are connected when a pair of nodes coexists in an image. To construct the knowledge graph, we use images from Matterport3D dataset (Chang et al., 2017).

4.1.3 Graph Encoder

We use a Graph Convolutional Network (GCN) (Kipf & Welling, 2017) to extract a graph feature from our knowledge graph. A GCN follows this layer-wise propagation rule below.

$$H^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}),$$

where $H^{(l)}$ is the representation of nodes at the layer l. $\tilde{A} = A + I_N$, and A is the adjacency matrix of the undirected graph G. I_N denotes the identity matrix, $\tilde{D}$ is the degree matrix of $\tilde{A}$, and $W^{(l)}$ are learned parameter weights at layer l. Let σ denote the activation function, which is ReLU in our implementation. As shown in Figure 2, our GCN takes joint embedding composed of word embeddings of the object or the region name and image embeddings from the vision encoder.

4.1.4 Vision Encoder and Location Predictor

Our vision encoder f_e^v encodes the visual observations, consisting of egocentric RGB and depth images from the agent's perspective. We used the pre-trained vision model described in section 4.1.1 as the vision encoder's backbone architecture.

The audio observation contains information about the relative distance and direction from the agent to the sounding object. Thus, we trained a location predictor f_{loc}^a to predict a location

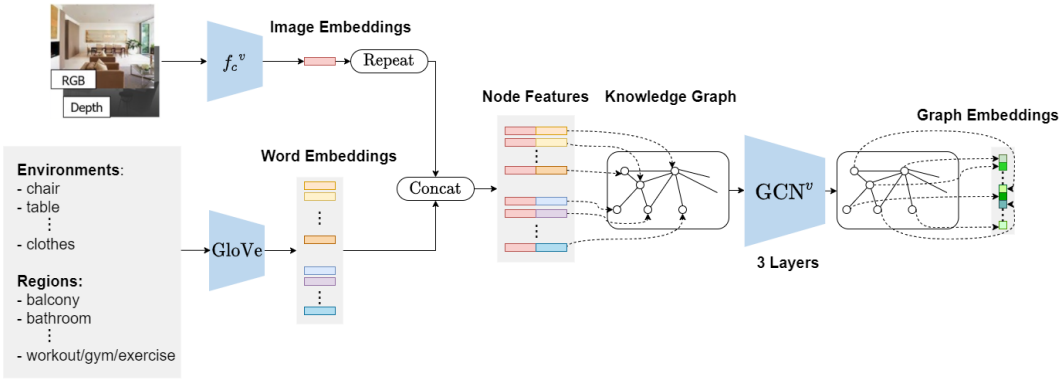

Figure 2: Graph Convolutional Networks. Each vertex denotes an object category or region category. The initial node features which are fed into the GCN are initialized with the joint embedding obtained by concatenating word embeddings of object or region name and image embeddings. Node information propagates through the three layers of the GCN, and the output of the GCN is graph embedding, which is stored as a memory in visoion networks.

$\hat{l}_t^a = (\Delta x, \Delta y)$ relative to the current pose p_t of the agent. Similar to the pre-trained audio model, location predictor model also aggregates the current prediction $\hat{l}_t^a$ with the previous prediction l_{t-1}^a, $l_t^a = f_\lambda(\hat{l}_t^a, l_{t-1}^a, \Delta p_t) = (1-\lambda)\hat{l}_t^a + \lambda f_p(l_{t-1}^a, \Delta p_t)$, where $f_p(\cdot)$ transforms the previous location prediction l_{t-1}^a based on the last pose change Δp_t, and λ is the weighting factor set to 0.5. The agent uses its latest estimate $l_t^a = f_p(l_{t-1}^a, \Delta p_t)$ when the acoustic event stops. We used the pre-trained audio model described in section 4.1.1 as the backbone architecture for training the location predictor.

4.1.5 Policy Network

We used attention-based transformer architecture for our reinforcement learning policy network, which stores observations in memory M. At each time step, our model encodes each visual observation, $e_t^v = f_e^v(v_t)$ and $e_t^{v-gcn} = GCN^v(f_c^v(v_t))$ to save in the memory. Our model also stores p, the agent's pose defined by its location and orientation (x, y, θ) with respect to its starting pose p_0 in the current episode, and a_{t-1}, the previous executed action, in the memory. Thus, the observation encoding vector stored in memory is $e_t^O = [e_t^v, e_t^{v-gcn}, p_t, a_{t-1}]$. The model stores these observations encoding up to time t in memory $M = \{e_t^O : i = max\{0, t - S_M\}, \ldots, t\}$, where S_M is the memory size.

The transformer uses the memory M stored so far in the episode and encodes these visual observation embeddings with a self-attention mechanism to compute the encoded memory $M_e = Encoder(M)$. Then, using the audio observation embeddings, a decoder network attends to all cells in the encoded memory M_e to calculate the state representation $s_t = Decoder(M_e, c_t^a, l_t^a)$. Using this attention mechanism, the agent captures long-term spatio-temporal associations between the acoustic-driven goal prediction and the visual observations. Moreover, our model preserves the most relevant information to reach the goal by conditioning visual-semantic embeddings stored in M_e on audio-semantic embeddings computed using current audio observation. The actor-critic network uses s_t to predict the value of the state and action distribution. Finally, the action sampler samples the next action a_t from this action distribution to select the agent's next action.

4.2 Learning and Optimisation

To train the vision classification model f_c^v, we collect a dataset using 85 Matterport3D houses, consisting of 82,828 images, each corresponding to a location and rotation angle in the SoundSpaces simulator. Each image has 128 x 128 resolution and 4 modalities: RGB image, depth image, object semantic image, and region semantic image. We filter out some images and only use 45,233 images to train our model (refer to appendix C for more details). We use the binary cross-entropy loss for optimizing the vision classification model and train it as a standard multi-label classifier.

To train the audio classification model f_c^a, we use the SoundSpaces simulator to generate 1.5M spectrograms using different source and receiver positions, each corresponding to a sounding object present in one of the 85 Matterport3D houses. One spectrogram corresponds to a sounding object, which could be present in multiple regions. For example, a sink can be present in the kitchen and bathroom regions. Thus, we treat detecting sounding objects as a multi-class classification problem and detecting regions in which that sounding object could be present as a multi-label classification problem. Furthermore, we optimize the audio classification model using cross-entropy loss for sounding object detection and binary cross-entropy loss for detecting regions.

Our vision classification model takes an RGB image as input, and the audio classification model takes 1 second sound clip represented as two 65 × 26 binaural spectrograms as input. We trained both vision classification and audio classification models using a ResNet-18 (He et al., 2015) architecture pre-trained on ImageNet to predict a score to 21 objects and 24 regions. These models are pre-trained before training the agent and kept frozen during policy training.

For training the location predictor f_{loc}^a to predict a relative location of the sounding object, we use the ResNet-18 architecture and initialize it with the weights of the pre-trained audio classification model. Location predictor is trained during policy training using the same experience collected for policy training. We optimize the location predictor using the mean squared error loss and update it with the same frequency as the policy network.

We train the policy network using the decentralized distributed proximal policy optimization (DD-PPO) (Wijmans et al., 2020b), which consists of a value network loss, policy network loss, and an entropy loss to encourage exploration (Schulman et al., 2017). We adapt the two-step training procedure proposed in Fang et al. (2019) for effectively training the vision networks (f_e^v, GCN^v). In the first step, the SMT policy is trained without attention by setting the memory size $s_M = 1$ and storing the latest observation embeddings. In the second step, the memory size is set to $s_M = 150$, and the parameters of the vision networks are frozen. Training SMT requires enormous computational power, and due to limited computational resources, we were not able to complete the second step of training the SMT policy. Thus, the results for our method and the SAVi baseline correspond to the policy after the 20,000 updates of the first training step. Moreover, the results for the rest of the baselines also correspond to the policy after 20,000 updates. We emphasize that this may not be a fair comparison because some policy converges sooner than other.

The input to the vision encoder f_e^v is 64×64 RGB, and depth images are cropped from the center. We optimise our model using Adam (Kingma & Ba, 2015) with a learning rate of 2.5×10^{-4} for the policy network and 1×10^{-3} for the pre-trained audio and vision networks using PyTorch (Paszke et al., 2019).

5 EXPERIMENTS

5.1 ENVIRONMENT

Simulator and Semantic Sounds. We use SoundSpaces (Chen et al., 2020), a visually- and acoustically-realistic simulation platform, to simulate an agent navigating in 3D house environments. The simulator renders sounds at any pair of source (sounding object) and receiver (agent) locations on a uniform grid of nodes spaced by 1 meter. While, SoundSpaces supports two real-world environment scans (Replica (Straub et al., 2019) and Matterport3D (Chang et al., 2017)), we used Matterport3D as it provides a larger number of houses and object-region semantics therein. We use the same 21 object categories as Chen et al. (2021a) for Matterport3D: chair, table, picture, cabinet, cushion, sofa, bed, chest of drawers, plant, sink, toilet, stool, towel, tv monitor, shower, bathtub, counter, fireplace, gym equipment, seating, and clothes. These object categories are visually present in the 24 regions (balcony, bathroom, bedroom, closet, dining room, entryway/foyer/lobby, familyroom/lounge, hallway, junk, kitchen, laundryroom/mudroom, living room, lounge, meetingroom/conferenceroom, office, other room, porch/terrace/deck, rec/game, spa/sauna, toilet, utilityroom/toolroom, and workout/gym/exercise) of the 85 Matterport3D houses. We use the publicly available sound clips from the experiment performed by Chen et al. (2021a), in which audio clips from `freesound.org` database were used. We generate sound by rendering the specific sound that semantically matches the object at the locations in Matterport3D houses. For example, the water-dropping sound will be associated with the sink in the kitchen.

Rewards and Episodes. The agent receives a sparse reward of $+10$ when it reaches the goal successfully, a dense reward of $+1$ for reducing the geodesic distance to the goal, and an equivalent negative reward for increasing it. To encourage trajectory efficiency, we also assign a negative reward of -0.01 per time step. To avoid simpler episodes, in which it is easy to reach goal (e.g., straight paths or short distance), we used 2 conditions while sampling episodes: 1) the ratio of geodesic distance to euclidean distance must be greater than 1.1; 2) the geodesic distance from the start location to the goal location must be greater than 4 meters. We sample 367,155 episodes for training and 1000 episodes for each of the testing settings.

Action space and sensors. There are 4 actions in the agent's action space: *MoveForward*, *TurnLeft*, *TurnRight*, and *Stop*. *MoveForward* changes the agent's current location to the node in front of it only if that node is reachable without collision. *Stop* can be used by the agent to report sounding objects and terminal the episode. The *TurnLeft*, *TurnRight*, and *Stop* actions can always be executed successfully. There are 4 sensory inputs: egocentric binaural sound (two-channel audio waveforms), RGB image, depth image, and the agent's current pose relative to the starting pose of the episode. The resolutions of the RGB and depth images are 128×128.

Episode specification and success criteria. An episode of semantic audio-visual embodied navigation task is defined by a house, a start location, and rotation angle of the agent, a goal location, a sounding object, and duration of the audio event. In each episode, the start location and rotation of the agent is randomly selected. For selecting the sounding object, an instance of an object category in the house is also chosen randomly. We define a set of viewpoints within 1 meter of the object's boundary for each sounding object. When the agent executes *Stop* action at any of these viewpoints, the episode will be successfully completed.

5.2 BASELINES

We compare the performance of our model with the following baselines:

1. **Random walk**, a baseline which uniformly samples one of the three navigation actions and executes *Stop* automatically when the target sounding object is reached within 1m radius.
2. **AudioGoal** (Chen et al., 2020), an end-to-end RL policy based on the PointGoal task (Wijmans et al., 2020a) based on a Seq2Seq mechanism which uses a GRU state encoder that leverages colour and depth images to navigate the unknown environments. In contrast to PointGoal which uses GPS sensing to guide the agent toward its goal, this baseline uses audio spectrograms.
3. **AudioObjectGoal** a Seq2Seq mechanism similar to (2) but the agent is also provided with the semantic label of the target object.
4. **SAVi** (Chen et al., 2021a), a transformer-based model that uses a goal descriptor network, which predicts both spatial and semantic properties of the target sounding object. It is the state-of-the-art deep reinforcement learning model for the semantic audio-visual embodied navigation task.
5. **`K-SAVEN`**, the model proposed in this paper.

5.3 EVALUATION METRICS AND RESULTS

We considered the following metrics for evaluating agent performance: 1) success rate: the proportion of episodes in which the agent stops exactly at one of the viewpoints of the sounding object on the grid; 2) success rate normalized by the inverse path length (SPL): a standardised metric (Anderson et al., 2018a) for capturing information about trajectory length-efficiency; 3) success rate normalized by the inverse number of actions (SNA) (Chen et al., 2021b): this metric penalises in-place rotation actions and collisions which do not lead to path changes, as a proxy for action-efficiency; 4) average distance to goal (DTG): the average distance between the agent and the goal when episodes are finished; 5) success when silent (SWS): the proportion of successful episodes when the agent reaches the target sounding object after the end of the acoustic event.

We analyse the generalisation ability of our method by evaluating it on unheard sounding objects. More specifically, we evaluate our model on the following testing settings: 1) test on seen houses with unheard sounding object categories as the navigation target; and 2) test on unseen houses with

Table 1: Results of baseline models and our proposed approach on the Semantic Audio-Visual Embodied Navigation (SAVEN) task, under SoundSpaces simulation (Chen et al., 2020).

	SEEN HOUSES, UNHEARD SOUNDS					UNSEEN HOUSES, UNHEARD SOUNDS				
Method	SR (↑)	SPL (↑)	SNA (↑)	DTG (↓)	SWS (↑)	SR (↑)	SPL (↑)	SNA (↑)	DTG (↓)	SWS (↑)
Random	6.75	1.58	0.67	16.86	6.75	6.94	2.04	0.87	15.20	6.64
AudioGoal (Chen et al., 2020)	11.47	**10.78**	7.52	14.06	3.62	12.49	11.63	7.86	14.16	3.83
AudioObjectGoal	11.47	10.67	**8.18**	14.73	2.69	15.00	**13.97**	**9.62**	13.06	4.53
SAVi (Chen et al., 2021a)	13.33	7.91	7.60	11.26	6.82	10.47	5.75	4.80	11.85	5.94
K-SAVEN (ours)	**16.63**	8.64	5.41	**9.42**	**14.57**	**15.71**	6.52	3.77	**10.49**	**13.90**

unheard sounding object categories. We randomly split the houses and sounding objects for training and testing. More specifically, we use 68 seen houses, 17 unseen houses, 16 heard sounding objects, and 5 unheard sounding objects. We average the results over 1,000 episodes for each testing setting.

Results. Table 1 shows the results of the experiments conducted on the baselines introduced on the previous section. The Random baseline exhibits the lowest performance, compared against all other baselines, due to the challenging nature of the task. Nonetheless, despite having no learning components, it is able to achieve around 7% success rate. The AudioGoal and AudioObjectGoal baselines perform comparably in all metrics and in both of the experimental setups. Both of the aforementioned learning-based approaches were trained on a single training stage for 6M steps, until convergence.

Conversely, for both SAVi and our model, K-SAVEN, we present the results of the pre-training stage of the models. These results show that our approach outperforms the baselines in most of the evaluation metrics. Both, SPL and SNA were the two metrics where our model was outperformed by the Seq2Seq baselines. We highlight that during the pre-training stage of K-SAVEN, the agent is encouraged to explore the environments. Thus, at this stage our agent performs a higher number of steps (175 on average) per episode, compared to the AudioGoal and AudioObjectGoal baselines (50). Moreover, our model also performs significantly better than random. This shows that our model enables the agent to reason about the semantics of objects and regions, and navigate to a completely novel and unheard sounding object more efficiently. Once we complete the entire training process, we expect to see a significant improvement in performance for our model.

6 DISCUSSION AND CONCLUSION

We introduce a framework for leveraging knowledge-enhanced scene priors, in the form of object and region semantics, for the semantic audio-visual navigation task. Notably, we show performance improvements over strong baselines in multiple unseen contexts, particularly in conditions where the agent needed to find novel sounding objects. We also provid a knowledge graph for training models, a curated visual dataset, and a new task definition–all guided towards developing and assessing model generalisation performance in unseen environments.

We recognise future improvements of our work, e.g., in the selection of the knowledge resource used for encouraging scene priors in the semantic audio-visual navigation task. We would consider constructing a knowledge resource that characterises sound-object relations (i.e., with descriptions of the sound that is generated by various objects), more befitting of pre-training the acoustic GCN stream. Furthermore, we can consider using scene priors as additive modules on frameworks in other tasks, particularly within the family of embodied multimodal planning. Finally, sounds are not merely a product of individual objects, but of different types of interactions (e.g., sitting, dropping, playing music) that often involve multiple objects and/or people. Therefore, in future work, we plan to incorporate such semantic knowledge about sounds, objects, and interactions in our knowledge graphs to further improve performance.

REFERENCES

Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Roshan Zamir. On evaluation of embodied navigation agents. *CoRR*, abs/1807.06757, 2018a. URL http://arxiv.org/abs/1807.06757.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3674–3683. Computer Vision Foundation / IEEE Computer Society, 2018b. doi: 10.1109/CVPR.2018.00387. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In *arXiv:2006.13171*, 2020.

John F. Canny. *The Complexity of Robot Motion Planning*. MIT Press, Cambridge, MA, USA, 1988. ISBN 0262031361.

Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pp. 667–676. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00081. URL https://doi.org/10.1109/3DV.2017.00081.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Russ R. Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020b.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological SLAM for visual navigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12872–12881. Computer Vision Foundation / IEEE, 2020c. doi: 10.1109/CVPR42600.2020.01289. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chaplot_Neural_Topological_SLAM_for_Visual_Navigation_CVPR_2020_paper.html.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pp. 17–36. Springer, 2020. doi: 10.1007/978-3-030-58539-6_2. URL https://doi.org/10.1007/978-3-030-58539-6_2.

Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15516–15525. Computer Vision Foundation / IEEE, 2021a. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.html.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *9th International*

Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021b. URL https://openreview.net/forum?id=cR91FAodFMe.

Heming Du, Xin Yu, and L. Zheng. Learning object relation graph and tentative policy for visual navigation. *ArXiv*, abs/2007.11018, 2020.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 538–547. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00063. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Fang_Scene_Memory_Transformer_for_Embodied_Agents_in_Long-Horizon_Tasks_CVPR_2019_paper.html.

Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *arXiv preprint arXiv:2106.13948*, 2021.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9701–9707. IEEE, 2020.

Daniel Gordon, Dieter Fox, and Ali Farhadi. What should I do now? marrying reinforcement learning and symbolic planning. *CoRR*, abs/1901.01492, 2019. URL http://arxiv.org/abs/1901.01492.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. Hierarchical cross-modal agent for robotics vision-and-language navigation. *CoRR*, abs/2104.10674, 2021. URL https://arxiv.org/abs/2104.10674.

Lydia E. Kavraki, Petr Svestka, Jean-Claude Latombe, and Mark H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robotics Autom.*, 12(4): 566–580, 1996. doi: 10.1109/70.508439. URL https://doi.org/10.1109/70.508439.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Sven Koenig and Maxim Likhachev. Real-time adaptive a*. In Hideyuki Nakashima, Michael P. Wellman, Gerhard Weiss, and Peter Stone (eds.), *5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan, May 8-12, 2006*, pp. 281–288. ACM, 2006. doi: 10.1145/1160633.1160682. URL https://doi.org/10.1145/1160633.1160682.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: an interactive 3d environment for visual AI. *CoRR*, abs/1712.05474, 2017. URL http://arxiv.org/abs/1712.05474.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv:1602.07332*, 2016a. URL https://arxiv.org/abs/1602.07332.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016b.

Steven M. Lavalle, James J. Kuffner, and Jr. Rapidly-exploring random trees: Progress and prospects. In *Algorithmic and Computational Robotics: New Directions*, pp. 293–308, 2000.

Yunlian Lv, Ning Xie, Yimin Shi, Zijiao Wang, and Heng Tao Shen. Improving target-driven visual navigation with attention on 3d spatial relationships, 2020.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pp. 22–32, 2019.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13507–13515, 2021.

Mahdi Kazemi Moghaddam, Qi Wu, Ehsan Abbasnejad, and Javen Qinfeng Shi. Optimistic agent: Accurate graph-based value estimation for more successful visual navigation, 2020.

Alessandro Oltramari, Jonathan Francis, Cory Henson, Kaixin Ma, and Ruwan Wickramarachchi. Neuro-symbolic architectures for context understanding. In *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, pp. 143–160. IOS Press, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL http://arxiv.org/abs/1912.01703.

Yiding Qiu, Anwesan Pal, and Henrik I. Christensen. Learning hierarchical relationships for object-goal navigation, 2020.

Homagni Saha, Fateme Fotouhif, Qisai Liu, and Soumik Sarkar. A modular vision language navigation and manipulation framework for long horizon compositional tasks in indoor environment. *CoRR*, abs/2101.07891, 2021. URL https://arxiv.org/abs/2101.07891.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, 2016. URL http://arxiv.org/abs/1612.03975.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The replica dataset: A digital replica of indoor spaces. *CoRR*, abs/1906.05797, 2019. URL http://arxiv.org/abs/1906.05797.

Varun Kumar Vijay, Abhinav Ganesh, Hanlin Tang, and Arjun Bansal. Generalization to novel objects using prior relational knowledge, 2019.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a. URL https://openreview.net/forum?id=H1gX8C4YPr.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=H1gX8C4YPr.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, Jan 2021. ISSN 2162-2388. doi: 10.1109/tnnls.2020.2978386. URL http://dx.doi.org/10.1109/TNNLS.2020.2978386.

Wei Yang, X. Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *ArXiv*, abs/1810.06543, 2019.

A Additional Knowledge Graph Details

Knowledge Graph Construction Our knowledge graph captures object-to-object, object-to-region, region-to-object and region-to-region relations. To compute these relations, we use the semantic labels of objects and regions in Matterport3D. The heuristic we use to find these relations is frequency-based. The main idea is to connect an object with another object if they frequently exist in different regions. Similarly, connect a region with another region if they both have similar objects placed in them.

B Additional Model Implementation Details

Hyperparameters. For all experiments, we implemented models using the `PyTorch` deep learning library, version 1.8.0. We directly utilised standard implementations of the Adam optimiser Kingma & Ba (2014), with a learning rate of 0.003. During training, we used a batch size of 256 for all implementations.

Computing hardware. For rendering the simulator and performing local agent verification and analysis, we used a single GPU machine, with the following CPU specifications: Intel(R) Core(TM) i5-4690K CPU @ 3.50GHz; 1 CPU, 4 physical cores per CPU, total of 4 logical CPU units. The machine includes a single GeForce GTX TITAN X GPU, with 12.2GB GPU memory. For generating multi-instance experimental results, we used a dual-GPU machine, with the following CPU specifications: Intel(R) Core(TM) i9-9920X CPU @ 3.50GHz; 1 CPU, 12 physical cores per CPU, total 24 logical CPU units. The machine includes two NVIDIA Titan RTX GPUs, each with 24GB GPU memory.

C Vision dataset Generation Details

To train the vision classification model f_c^v, which given an RGB image predicts a score for objects and regions, we collect a vision dataset using the SoundSpaces simulator as described in Section 4.2. Initially, we collected 82,828 images across 85 Matterport3D houses, which is the maximum number of images possible as there are a total of 20,707 nodes and 4 rotation angles in SoundSpaces. However, we faced the following challenges with the scans and semantic labeling in the Matterport3D: 1) objects are not clearly visible because of glitches in scans (see RGB image in Figure 3); 2) object and region semantic labels are improper (see object and region semantic images in Figure 3); 3) objects are not semantically placed (see Figure 4).

To address these challenges, we filtered some images and only used 45,233 images to train our vision classification model f_c^v. We use the following filtration criteria: 1) Filter out an image in which 75% of the pixels or more are black (zero value); 2) There are 42 objects in Matterport3D, and we are interested in only 21 objects in our experiments, so we filter out an image if it does not contain any of

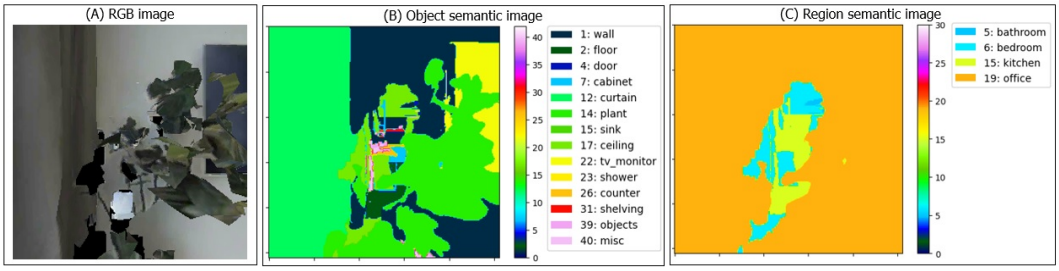

Figure 3: Examples of issues with the scans and semantic labeling in the Matterport3D. The images shown correspond to the scene ID aayBHfsNo7d and node: (93, 270). (A) shows an RGB image example in which the objects are not clearly visible due to glitches in the scan. (B) shows the semantic labels of 14 objects, and (C) shows the semantic labels of 4 regions; however, these objects and regions are not clearly visible in the corresponding RGB image.

Figure 4: Examples of unusual semantically placed objects in scene ID 2n8kARJN3HM of Matterport3D. In the image show, a bathtub is placed in the living room, and chairs are kept on the top of a table, which is unusual placement of these objects.

those 21 objects; 3) Filter out an image if the most frequent object is taking less than 3% of the total pixels in the image; 4) Filter some of the semantic labels of an image based on a threshold (0.18 for object and 0.2 for region). First, for each semantic label in the image, we compute the ratio of its proportion of the pixels to the proportion of the most frequent semantic label in that image. Then, semantic labels with ratios less than the threshold are filtered out.