ForesightNav: Learning Scene Imagination for Efficient Exploration

Supplementary Material

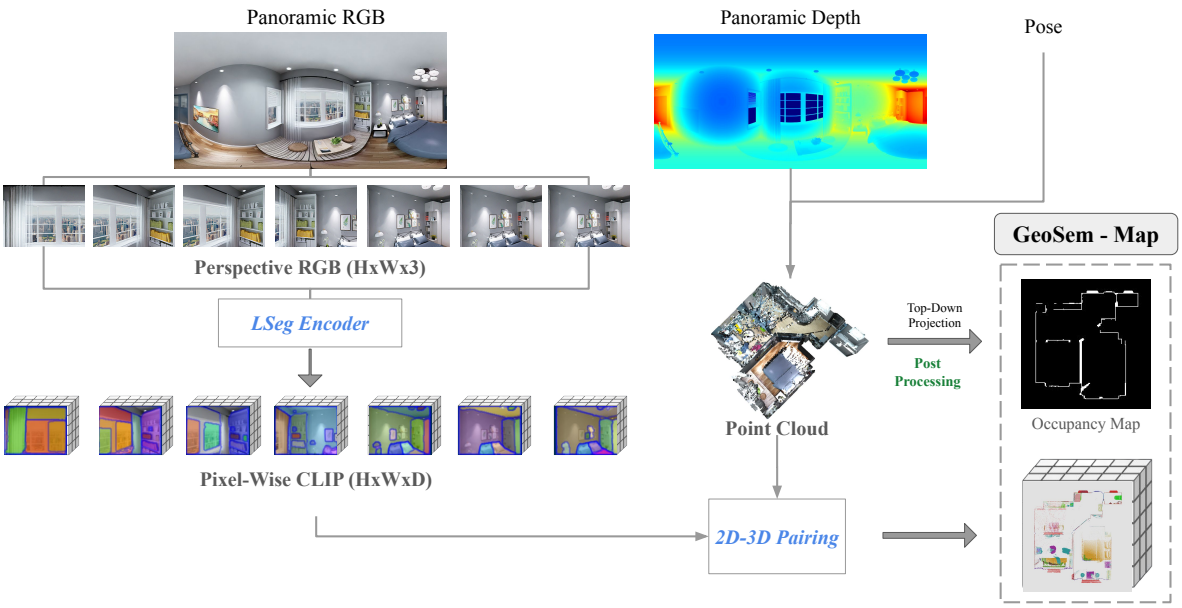

Figure 8. GeoSem Map generation for the Structured3D Dataset (Equirectangular RGBD)

7. GeoSem Maps from Equirectangular RGBD + Pose

The Structured3D dataset [41] provides richly annotated 3D indoor scenes with panoramic RGBD observations. It contains 3,500 scenes in total, split into 3,000 for training and 500 for validation. Each scene S_i is composed of multiple rooms, and for every room R_{ij}, the dataset offers a set of equirectangular RGB images $I_{\text{equirect}_{ij}} \in \mathbb{R}^{H\times W\times 3}$, corresponding equirectangular dense depth maps $D_{\text{equirect}_{ij}} \in \mathbb{R}^{H\times W\times 1}$, and global camera poses $T_{ij} \in SE(3)$ for each panoramic viewpoint.

In this section, we describe the construction of *GeoSem Maps* for each scene, which serve as groundtruth supervision for the imagination module discussed in Section 3.2. A representative Figure 8 depicts the overview of the method. Directly using the equirectangular images I_{equirect} as input to the pre-trained LSeg encoder leads to degraded CLIP embeddings due to the geometric distortions introduced by the panoramic format. To mitigate this, we instead apply a perspective projection strategy, simulating standard pinhole camera views to generate high-quality, per-pixel CLIP embeddings.

7.1. Perspective Projection for GeoSem Map Construction

We utilize a perspective camera model with a field of view (FOV) of 90 degrees. The viewing directions are strategically chosen so that the complete room is covered: horizontal view directions $\Theta = [-180°, -90°, 0°, 90°]$ and vertical view directions $\Phi = [-45°, 0°, 45°]$. Each combination of Θ and Φ results in a distinct perspective image capturing the room from different orientations i.e. 12 viewing directions. Given an equirectangular image $I_{\text{equirect}_{ij}}$ of scene S_i and room R_{ij}, perspective projection gives us 12 images $I_{\text{persp}_{ijk}}$ where $k \in [1, 12]$. With perspective projection, we loose the pixel-wise depth association between I_{equirect} and D_{equirect}. To finally construct a GeoSem Map as described in 3.1, we need pixel-wise CLIP embeddings and depth, along with camera pose for top-down projection onto a BEV map.

7.2. 2D-3D Correspondence Using Indexing Approach

Each pixel (u, v) in $D_{\text{equirect}_{ij}}$ represents a depth value d_{uv}, which can be projected into a local 3D point $\mathbf{p}_{uv}^{\text{local}} \in \mathbb{R}^3$ using the intrinsic parameters of the equirectangular projection. These local 3D points are then transformed into the global coordinate frame using the camera pose:

$$\mathbf{p}_{uv}^{\text{global}} = T_{ij} \cdot \mathbf{p}_{uv}^{\text{local}}$$

A global point cloud is constructed for the scene by aggregating all such transformed 3D points across rooms:

$$\mathcal{P}_i = \bigcup_j \{\mathbf{p}_{uv}^{\text{global}} \mid (u,v) \in D_{\text{equirect}_{ij}}\}$$

To retain the pixel-to-point correspondence for associating the CLIP embeddings later, an indexing scheme is used:

- **Base Indexing:** Each panoramic image $I_{\text{equirect}_{ij}}$ is assigned a unique base index B_{ij}.
- **Pixel Indexing:** We create the index map $\texttt{index}_{ij} \in \mathbb{R}^{H \times W \times 1}$, where each pixel (u,v) receives a unique global index within the scene:

$$\texttt{index}_{ij}(u,v) = B_{ij} + (u \cdot W + v)$$

This ensures that every pixel across all panoramas in a scene has a globally unique identifier.

Perspective Projection and Index Transformation: When we convert $I_{\text{equirect}_{ij}}$ into a perspective image $I_{\text{persp}_{ijk}}$, we similarly project the index map $\texttt{index}_{ij}$ to obtain a corresponding *index perspective image* ϕ_{ijk} which allows us to define the mapping:

$$\phi_{ijk} : \text{Pixels in } I_{\text{persp}_{ijk}} \rightarrow \text{Point IDs in } \mathcal{P}_i$$

$$\phi_{ijk}(u',v') = \texttt{index}_{ij}(u,v)$$

This enables each pixel in the perspective image to retain the identity of its original 3D point, thereby preserving the connection between CLIP embeddings (obtained from LSeg) and their corresponding 3D positions in the scene.

$$\text{LSeg}(I_{\text{persp}_{ijk}})(u',v') \rightarrow \mathcal{P}_i[\phi_{ijk}(u',v')]$$

7.3. CLIP Embedding Generation and Mapping

With the perspective images and corresponding depth information associated using the index images, we employ the LSeg encoder to derive pixel-wise CLIP embeddings for each perspective image.

The subsequent steps involve aggregating these CLIP embeddings into a grid representation $\mathcal{M}$, where each grid cell encapsulates semantic information from the underlying 3D points and their associated CLIP embeddings. The construction of the GeoSem Map $\mathcal{M}$ proceeds as outlined in Section 3.1.

7.4. Handling Spurious Observations

From the Structured3D annotations, the groundtruth floor plan is known. However, to better simulate real-world conditions during training, we generate the groundtruth occupancy maps using a top-down projection of the global point-cloud instead of using the dataset floor plan annotation. Due to imperfections in the point cloud, the resulting occupancy maps may contain small holes in the walls. These can cause "leaks" during agent simulation (described in Section 4) where depth rays may erroneously pass through walls. This leads to artifacts in the occupancy map - specifically, observed regions outside the intended scene boundaries which would not occur in a realistic setting. To address this, we generate an interior-exterior mask $\mathbf{E} \in \{0,1\}^{M \times M}$ from the Structured3D annotations for room polygons, marking all points in the exterior as 0.5 (unobserved) in the occupancy map P:

$$\mathbf{P} = \mathbf{P} \odot \mathbf{E} + 0.5(1 - \mathbf{E}) \tag{14}$$