

# Active Neural Mapping at Scale

Zijia Kuang<sup>1</sup>, Zike Yan<sup>1,†</sup>, Hao Zhao<sup>1</sup>, Guyue Zhou<sup>1</sup>, and Hongbin Zha<sup>2</sup>

**Abstract**—We introduce a NeRF-based active mapping system that enables efficient and robust exploration of large-scale indoor environments. The key to our approach is the extraction of a generalized Voronoi graph (GVG) from the continually updated neural map, leading to the synergistic integration of scene geometry, appearance, topology, and uncertainty. Anchoring uncertain areas induced by the neural map to the vertices of GVG allows the exploration to undergo adaptive granularity along a safe path that traverses unknown areas efficiently. Harnessing a modern hybrid NeRF representation, the proposed system achieves competitive results in terms of reconstruction accuracy, coverage completeness, and exploration efficiency even when scaling up to large indoor environments. Extensive results at different scales validate the efficacy of the proposed system.

## I. INTRODUCTION

Accurate modeling of the surrounding environment for an embodied agent is of great importance towards spatial intelligence. Recent pivotal advances in the relevant fields are the developments of implicit neural representations (INRs) [1], [2], [3], [4]. Scene reconstruction is formulated as a coordinate-based low-dimensional regression problem and solved through gradient-based optimization, where a compact model can be attained to recover accurate scene geometry [5], appearance [2], [6], and semantics [7], [8]. Nevertheless, the inherent under-determined nature as an inverse problem leads to susceptibility to artifacts, as the continuous scene representation is inferred from incomplete observations in an unknown environment.

Note that the ill-posed issue not only applies to the differentiable optimization of INRs, but also to the conventional data fusion paradigms based on discretization-based representations such as volume grids and meshes. Insufficient observations lead to incompleteness and thus require autonomous exploration and reconstruction of the environment, or namely active mapping, an area that has been studied for decades [10]. While the problem is well-formulated by approaching the surface frontiers [11] or by selecting the next-best-view samples through certain criteria [12], we aim to address in this work the following question: *Is information within a neural map sufficient for fast and thorough exploration in an unknown indoor environment?*

As defined in [11], exploration is “the act of moving through an unknown environment while building a map for subsequent navigation”. That is to say, the neural map should be updated on-the-fly, quantify uncertainty beyond

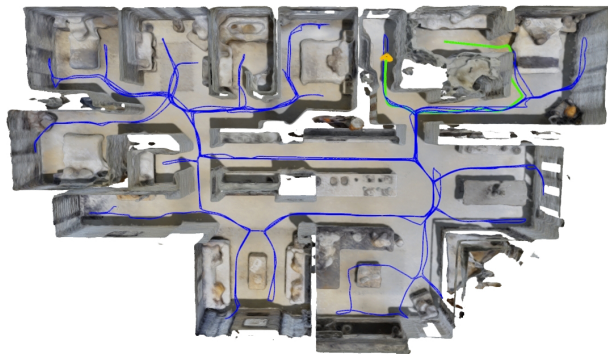


Fig. 1: The reconstructed mesh through autonomous exploration with a continually-learned neural map (MP3D-zsNo4 [9] with 23 rooms). The synergistic integration of geometry, appearance, topology, and uncertainty information within the neural map leads to complete and accurate reconstruction of large-scale environments.

past observations, and provide a collision-free path to traverse the environment. This is formulated in [13], namely “active neural mapping”, in a greedy fashion. Implicit neural representations have been challenged regarding the slow convergence, the lack of interpretability, and the absence of explicit structural information. While recent advances in implicit neural representation allow efficient online reconstruction [14], [15] and reasonable uncertainty quantification [13], [16], we argue that the lack of explicit structural information prevents INRs from interpreting granularity at different levels. This is one typical gap between neural and symbolic representations and hinders efficient exploration with NeRF-wise representations.

In this paper, we propose to extract a generalized Voronoi graph (GVG) from the neural map to organize the information at different levels of detail in a structured manner. Edges of GVG delineate the spatial partitions of surface points according to Voronoi tessellation, where the inherent sparsity and the focus on maximizing clearance allow efficient and safe path planning for autonomous exploration (see Fig. 1). As vertices of GVG symbolize the common boundaries of multiple sub-areas, they encapsulate high geometry complexity and serve as critical decision points. Therefore, we anchor the fine-grained details from the neural map to Voronoi vertices to balance between efficient exploration and information utilization. The exploration can then be conducted under adaptive granularity through information-guided traversal of

<sup>†</sup> Corresponding author. [yanzike@air.tsinghua.edu.cn](mailto:yanzike@air.tsinghua.edu.cn)

<sup>1</sup> Zijia Kuang, Zike Yan, Hao Zhao, and Guyue Zhou are with the Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China.

<sup>2</sup> Hongbin Zha is with the School of Intelligence Science and Technology, Peking University, Beijing, China

Voronoi vertices recursively.

- We introduce a NeRF-based active mapping system that can explore large-scale indoor environments with up to 20+ rooms comprehensively.
- We extract structured information within the neural map using a generalized Voronoi graph to take both accessible areas of interest and visible areas of interest into account.
- We maintain a hierarchical framework to balance between information utilization and exploration efficiency with competitive performance.

## II. RELATED WORK

*a) Autonomous exploration and reconstruction:* Autonomous exploration and reconstruction reduce the uncertainty of the map by finding the areas to be explored iteratively. Frontier-based methods maintain the boundary between explored and unvisited space [11] and gradually expand the coverage. This is usually achieved through mapping, frontier detection, and selection, where different strategies are made to extend the method to multi-robot scenario [17], [18] and 3D space [19], [20] with better completeness and accuracy. Another line of research tends to find the next-best-view in a receding horizon to maximize the target objective [12]. As a sampling-based method, attempts are made to determine the per-sample information gain [21], [22] and connect these samples through feasible paths.

There are also hybrid works [23], [24] that trade-off between local accuracy and global completeness for efficient and accurate exploration. Recent methods focus on reward-driven learning policy instead of heuristic designs. This is usually achieved in a modular [25], [26] or end-to-end [27], [28] fashion. The learned prior can also predict occupancy in unseen areas [29] to accelerate the exploration. We, on the other hand, exploit the expressive representation power of neural maps to guide agent movement using fail-safe graph-based planning strategy.

*b) NeRF-based SLAM:* The differentiable rendering fashion through a compact but expressive neural network [2] has been commonly adopted as one promising method for 3D reconstruction. The method is later extended to continual learning based solutions [30], [31] to handle constant distribution shifts from sequential observations. Keyframes are restored as replayed buffers to constantly constrain the optimization. The computational efficiency and global accuracy are further enhanced through different neural representations [32], [15], [14] and global optimization [33].

One critical issue for NeRF-based scene reconstruction is the ill-posed nature as an inverse problem [1]. Insufficient observations will lead to artifacts and prediction errors. Different methods tackle the problem in an offline [34] or online [35] setting. Recent studies [13], [36] extend the INR-based active mapping problem to an online scene-scale level in a greedy fashion. In this paper, we further exploit the structure information within the neural map to accelerate exploration with promising reconstruction accuracy.

## III. ACTIVE NEURAL MAPPING WITH NeRF

We target an active neural mapping [13] problem, where neural map parameters  $\theta^t$  and the action control  $\mathbf{a}^t$  are optimized recursively to best explore and reconstruct an environment unknown *a priori*. The problem can be solved by finding a sequence of accessible poses  $\{\mathbf{v}^t\}_{1:n} \in \mathbf{SE}(2)$ , such that the neural map  $f(\mathbf{x}; \theta) : \mathcal{X} \rightarrow \mathcal{Y}$  can be continually updated given new observations  $z(\mathbf{v}^t) \subset \mathcal{D}$  to gradually minimize the reconstruction error  $\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}(\mathcal{L}(\mathbf{x}, \mathbf{y}; \theta))$ . To maintain the high-fidelity geometry and appearance of the scene, we optimize an implicit representation  $f(\mathbf{x}; \theta) \rightarrow (\mathbf{c}, s)$  that maps 3D spatial coordinates to color  $\mathbf{c}$  and truncated signed distance (TSDF) values  $s$  [14]. The compact representation allows efficient query of coordinate values through a single forward pass. Depth and color images can be synthesized through volume rendering [31], [32], [14] as the weighted sum of all samples  $\mathbf{p}_i = \mathbf{o} + d_i \mathbf{r}, i \in \{1, \dots, N\}$  along a ray  $\mathbf{r}(u, v)$ :

$$\hat{\mathbf{c}}[u, v] = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i \mathbf{c}_i, \hat{d}[u, v] = \frac{1}{\sum_{i=1}^N w_i} \sum_{i=1}^N w_i d_i, \quad (1)$$

where the weight  $w_i$  is the multiplication of two Sigmoid functions with a truncation threshold  $\lambda_{tr}$  as:

$$w_i = \sigma\left(\frac{s_i}{\lambda_{tr}}\right) \sigma\left(-\frac{s_i}{\lambda_{tr}}\right). \quad (2)$$

Though the neural map provides continuous and fine-grained details of the environment, enumerating the entire cascade of decisions  $\{\mathbf{a}\}_{1:n}$  in the working space without *a priori* is non-trivial. Central to this work is a structured representation  $\{\mathcal{V}, \mathcal{E}\}$  that manages information within the neural map  $f(\mathbf{x}; \theta)$  and enforces efficient and thorough exploration of unknown environments.

### A. Structuring NeRF into a navigational map

To capture the essential navigational information within the network, we adopt Voronoi tessellation to divide the working space into sub-regions and capture the topology of the partially-observed accessible regions. The Voronoi graph itself is a roadmap of the free space [37], where edges  $\mathcal{E}$  define a set of accessible and connected waypoints that are equidistant to the scene surface, and vertices  $\mathcal{V}$  are the meet points that the edges terminate at. As all surface points can be assigned to the sparse GVG edges according to Voronoi tessellation, the graph structures the information within a neural map into a compact and sparse form.

Denote a distance function from coordinate  $\mathbf{x}$  to a surface point  $\mathbf{p}_i$  as  $d_i(\mathbf{x}) = \text{dist}(\mathbf{p}_i, \mathbf{x})$ . The Voronoi tessellation defines the sub-region  $S_i$  affected by a surface point  $\mathbf{p}_i$  as the set of points in the space that are closer to  $\mathbf{p}_i$  than to any other surface points as:

$$S_i = \{\mathbf{x} | d_i(\mathbf{x}) < d_j(\mathbf{x}) \text{ for all } j \neq i\}. \quad (3)$$

Note that the surface points  $\{\mathbf{p}_i\}_{1:m}$  can be efficiently

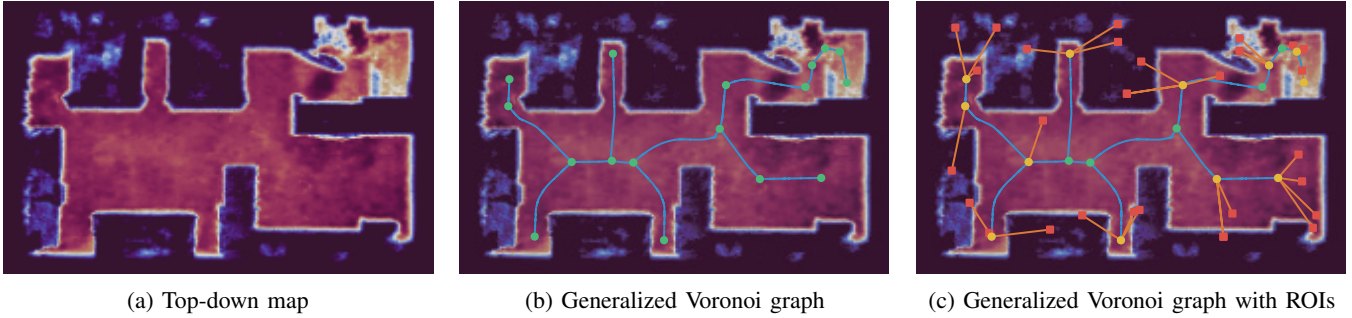


Fig. 2: The structuring process through Voronoi tessellation. The Generalized Voronoi graph (GVG) (b) is extracted as equidistant samples to the zero-crossing surfaces (a). The visible regions of interest (ROIs) guided by perturbed map parameters are anchored to neighboring vertices (c).

queried from the neural map as zero-crossings<sup>1</sup>. A Voronoi graph can be extracted [37] by first finding points equidistant to at least two surface points as edges, and then forming vertices as the intersections of multiple Voronoi edges. This can be done directly using the Voronoi tessellation by examining the edges of the polygons in the tessellation. We remove the vertices with negative distance values through forward passes. As illustrated in Fig. 2b, the Voronoi graph derives explicit topology of the partially observed free space that greatly simplifies the planning problem with a sparse and discrete structure.

### B. Identifying the region of interests

Given the topological structure of the accessible area, a path can be efficiently generated through graph-search algorithms once a target location is decided. With the partially observed topology and scene radiance from the instant Voronoi graph  $\{\mathcal{V}, \mathcal{E}\}^t$  and neural map  $f(\mathbf{x}; \theta^t)$ , we can take both accessible and visible region of interests into account for thorough exploration and accurate reconstruction.

To evaluate the reconstruction quality at a fine-grained level, we need to conduct uncertainty quantification from the continually learned neural map. Recent study [13] indicates that loss landscapes are divergent when evaluating prediction error at different areas: the areas in the working space that lack constraints would lead to unstable minima, where small perturbations on the network parameters would result in evident prediction variation; the areas with constant supervision during the continual learning process would lead to a flat basin that is robust against perturbation. Note that the adopted network simultaneously outputs the geometry  $s$  and the appearance  $c$  of the scene. Perturbing different decoders leads to different selections of interest. The test-time reparameterization would not affect the performance of the adopted backbone and is amenable to new advances. We perturb the parameters of both the geometry and appearance decoder of the neural field with Gaussian noises and quantify the prediction variances given zero-crossing coordinates

through multiple passes, where the surface points with higher variances will be considered as the visible area of interest in terms of both geometry and appearance:

$$\mathbf{x} = \arg \max \mathbb{V}_{\hat{\theta} \sim N(\theta^t, b^2 I)} [f(\mathbf{x}; \hat{\theta})]. \quad (4)$$

The visible region of interest inferred by the neural map is inherently dense and delicate. Geometric and photometric details would be frequently highlighted due to insufficient observations or training steps. Moreover, the target regions induced by the neural map may be non-traversable. Therefore, as illustrated in Fig. 2c, we cluster the visible regions of interest according to their geometric and topological information and anchor them to the near Voronoi vertices that serve as accessible regions of interest. We further depress the accessible region of interest if the Voronoi vertices fall into previously visited areas. Consequently, Voronoi vertices that are anchored with visible regions of interest or belong to non-visited areas would be activated as the target position candidates. This ensures a safe and traversable path recursively, where fine-grained observations can be achieved by adjusting the viewing angle towards the visible regions of interest once arriving at the selected vertex.

### C. Graph search for efficient planning

As the informative cues within the neural map are anchored at the sparse Voronoi vertices and edges, efficient action decisions can be made given the agent pose and the selected Voronoi vertex by treating the problem as a single-source shortest path planning. Given the explicit topology of the free space, we employ the Dijkstra algorithm [38] to find the optimal path in the weighted graph, where the weights are the Euclidean distance between two adjacent vertices. The path can be found through an iterative process that involves selecting the neighboring vertex with a minimal cumulative distance from the current position. Replanning is conducted every time the agent approaches the target frustum. The sparse nature of the graph guarantees efficient planning in milliseconds.

## IV. IMPLEMENTATIONS FOR LARGE-SCALE SCENES

Given the structured information, the complexity of areas to be visited greatly reduces to the sparse set of vertices

<sup>1</sup>A top-down map is created in the horizontal plane as illustrated in Fig. 2a. For each grid point in the top-down map, we take the minimum value of the queried SDF given samples along the vertical direction.

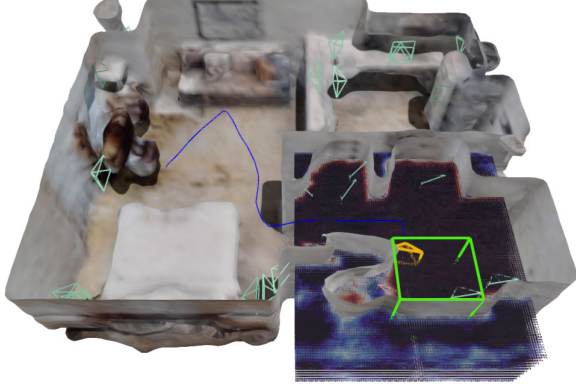


Fig. 3: The fine-grained local horizon (the dark area near the frustum that indicates the nearby distance values) guarantees safe exploration, while the rest of the map maintains the global regions of interest coarsely.

within the Voronoi graph. Unlike the greedy strategy conducted in [13] that is susceptible to local minima, the Voronoi graph assures completeness theoretically under certain conditions [39] and simplifies the search space to a graph structure without intersecting any obstacles. Despite the merits, the topology map based planning and the neural map based optimization need to be carefully designed to adapt to large-scale environments with satisfying efficiency and completeness.

#### A. Hierarchical framework with adaptive granularity

Voronoi graph based planning, while advantageous for many reasons, also encounters specific drawbacks when deployed in large-scale complex scenes. As proved by [40], [41], given that the vertex on the graph has a minimum degree of three, the undirected graph satisfies:

$$2|\mathcal{E}| = \sum_{v \in \mathcal{V}} \deg(v) \geq 3|\mathcal{V}|. \quad (5)$$

Meanwhile, viewing Voronoi graph as a planar graph with the number of faces  $l$  the same as the number of surface samples, Euler's formula leads to:

$$|\mathcal{V}| - |\mathcal{E}| + l = 2. \quad (6)$$

Therefore, the graph grows linearly with the number of surface samples:

$$|\mathcal{V}| \leq 3l, |\mathcal{E}| \leq 2l. \quad (7)$$

Besides the extraction of the Voronoi graph, the graph-based search algorithm has a runtime complexity of  $\mathcal{O}(|\mathcal{E}| + |\mathcal{V}| \log |\mathcal{V}|)$ , leading to the complexity of  $\mathcal{O}(l \log l)$ . When encountering a scene with complex surface geometry, the rising number of zero-crossings will greatly increase the computational cost of the planning. To this end, we maintain

a fixed resolution of the local horizon to extract precise zero-crossings, where coarse-level zero-crossings are extracted outside the local horizon. As illustrated in Fig. 3, the computational cost is upper-bounded while guaranteeing a safe path during exploration. This strategy greatly reduces the chances that the agent gets stuck by thin objects or near narrow zones while preserving promising efficiency. We prioritize anchored Voronoi vertices within the local horizon to avoid sudden maneuvers, and rotate toward the anchored regions of interest once arriving at the target position. Consequently, the agent tends to march quickly toward the selected Voronoi vertex and stay cautious at the critical decision points. The adaptive granularity for actively searching the informative observations achieves a nice balance between exploration efficiency and reconstruction accuracy.

#### B. Hybrid neural representation

The convergence of the neural representation greatly affects the exploration efficiency. As shown in [13], continual learning of a single multilayer perceptron (MLP) suffers from slow convergence, where the uncertainty of the neural map will guide the agent back to the visited areas with complex geometry details. In this paper, we adopt the joint coordinate and parametric encoding similar to Co-SLAM [14] to best tradeoff between fast convergence and accurate modeling. For scene geometry, the coordinate is encoded by the combination of a hash-encoded multi-resolution feature grid [3]  $\{\mathcal{F}_\alpha^l(\mathbf{x})\}_{l=1}^L$  and a One-blob encoding  $\gamma(\mathbf{x})$  [42], and decoded into the predicted TSDF value  $s$  and a feature  $\mathbf{h}$ . For scene radiance, the color is decoded given the concatenated features of  $\mathbf{h}$  and  $\gamma(\mathbf{x})$ :

$$f_\tau(\gamma(\mathbf{x}), \{\mathcal{F}_\alpha^l(\mathbf{x})\}_{l=1}^L) = \{\mathbf{h}, s\}, f_\phi(\gamma(\mathbf{x}), \mathbf{h}) = \mathbf{c}. \quad (8)$$

Denote the learnable parameters as  $\boldsymbol{\theta} = \{\alpha, \phi, \tau\}$ , we can perturb the shallow decoder of  $\gamma$  and  $\phi$  to quantify the uncertainty regarding the scene geometry and appearance and adjust the planning strategy accordingly. The training of networks takes a similar strategy with [14] to first penalize  $l_2$  errors of RGB and depth rendering:

$$\mathcal{L}_{rgb} = \frac{1}{N} \sum_{n=1}^N (\hat{c}_n - c_n)^2, \mathcal{L}_d = \frac{1}{|R_d|} \sum_{r \in R_d} (\hat{d}_r - D[u, v])^2, \quad (9)$$

where the SDF values within and without the truncated areas are penalized as either the depth values or the truncated distance  $tr$ :

$$\mathcal{L}_{sdf} = \frac{1}{|R_d|} \sum_{r \in R_d} \frac{1}{|S_r^{tr}|} \sum_{p \in S_r^{tr}} (s_p - (D[u, v] - d))^2, \quad (10)$$

$$\mathcal{L}_{free} = \frac{1}{|R_d|} \sum_{r \in R_d} \frac{1}{|S_r^{free}|} \sum_{p \in S_r^{free}} (s_p - tr)^2, \quad (11)$$

and a smoothness term is performed to regularize the grid feature as:

$$\mathcal{L}_{smooth} = \frac{1}{|\mathcal{G}|} \sum_{p \in \mathcal{G}} \Delta_x^2 + \Delta_y^2 + \Delta_z^2, \quad (12)$$

TABLE I: Comparison against relevant methods regarding the completeness ( $\% \uparrow / cm \downarrow$ ) of actively-captured observations.

	Random	FBE [11]	UPEN [43]	ANM [13]	Ours
<b>Gibson-Cantwell</b>	24.43/59.59	40.93/37.03	39.42/42.12	61.36/17.67	<b>85.09/7.25</b>
<b>Gibson-Denmark</b>	27.83/50.42	70.28/12.40	66.41/17.34	85.86/3.78	<b>91.85/2.00</b>
<b>Gibson-Eastville</b>	14.32/72.39	58.49/24.08	51.51/28.16	74.21/11.36	<b>87.66/6.82</b>
<b>Gibson-Elmira</b>	66.29/11.63	72.69/10.40	82.14/5.35	91.65/2.57	<b>95.42/1.40</b>
<b>Gibson-Eudora</b>	53.89/23.24	76.65/8.11	75.74/9.18	90.12/2.27	<b>93.94/1.49</b>
<b>Gibson-Greigsville</b>	75.44/6.97	90.34/2.62	73.72/16.34	92.47/1.78	<b>98.65/0.64</b>
<b>Gibson-Pablo</b>	46.87/34.70	76.06/6.38	54.16/31.81	72.88/9.96	<b>88.03/3.07</b>
<b>Gibson-Ribera</b>	44.29/33.27	79.26/6.53	81.21/5.74	88.62/4.13	<b>95.69/1.15</b>
<b>Gibson-Swarmville</b>	58.81/18.10	55.46/22.19	45.43/33.78	66.86/13.43	<b>92.58/1.68</b>
<b>Gibson-mean</b>	45.80/34.48	68.30/14.42	63.30/21.09	80.45/7.44	<b>92.10/2.83</b>
<b>MP3D-GdvgF</b>	68.45/11.67	81.78/5.48	82.39/5.14	80.99/5.69	<b>91.05/3.93</b>
<b>MP3D-gZ6f7</b>	29.81/46.48	81.01/7.06	82.96/6.14	80.68/7.43	<b>90.83/3.12</b>
<b>MP3D-pLe4w</b>	32.92/30.79	66.09/12.78	66.76/11.82	76.41/8.03	<b>96.20/1.68</b>
<b>MP3D-YmJkq</b>	50.26/24.61	68.32/11.85	60.47/15.77	79.35/8.46	<b>80.87/7.83</b>
<b>MP3D-mean</b>	45.36/28.39	74.30/9.29	75.56/9.72	79.36/7.40	<b>89.74/4.14</b>

where  $\Delta x, y, z = \mathcal{F}_\alpha(\mathbf{p} + \epsilon_{x,y,z}) - \mathcal{F}_\alpha(\mathbf{p})$  refers to the featuremetric distances between adjacent samples on the hash-grid.

## V. EXPERIMENTS

### A. Experimental setup

The experiments are conducted on a desktop PC with an Intel Core i9-10850K (10 cores @ 5.2 GHz), 64GB of RAM, and a single NVIDIA GeForce RTX 3090. We utilize the Habitat simulator [44] with the Gibson [45] and Matterport3D datasets [9] for quantitative and qualitative evaluation. The agent observes posed RGB-D image sequence at a resolution of  $256 \times 256$  and outputs discrete actions recursively, where the action space consists of MOVE\_FORWARD by  $6.5cm$ , TURN\_LEFT and TURN\_RIGHT by  $10^\circ$ , and STOP. The camera is set at a height of  $1.25m$  above the floor with a  $90^\circ$  field of view vertically and horizontally.

### B. Comparisons against other methods

We first follow the experimental setting of active neural mapping [13] to evaluate the exploration performance on 13 diverse scenes. The quantitative evaluation is conducted using completion ratio ( $\%$ )<sup>2</sup> and completion ( $cm$ ) metrics. As presented in Tab. I, the proposed active mapping system consistently outperforms the relevant methods of FBE [11], UPEN [43], and ANM [13]. The incorporation of the topology guarantees thorough exploration with promising reconstruction results (see 4). The system runs at 7-9HZ on average given different scales of scenes while achieving over 90% of completeness with fine-grained details of geometry and appearance. Nevertheless, as illustrated in Fig. 5, the

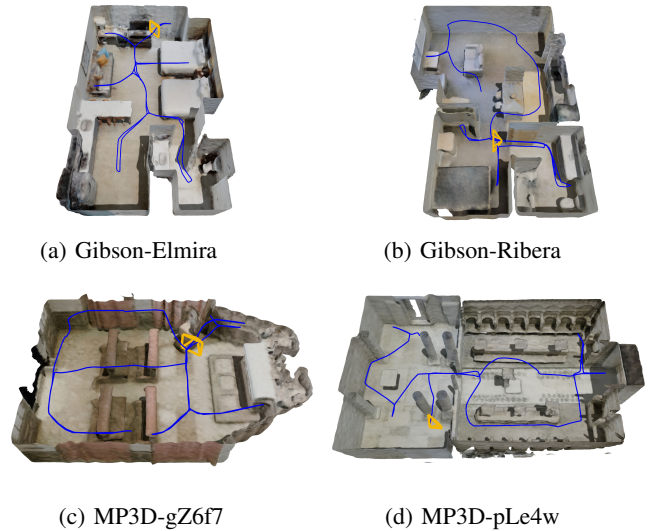


Fig. 4: The reconstruction results of small scenes through active exploration.

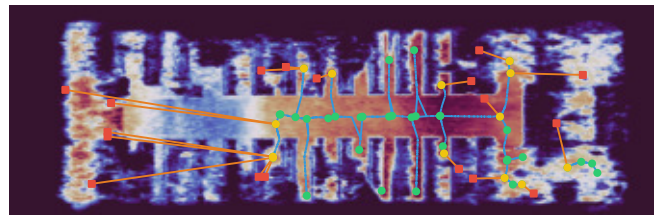
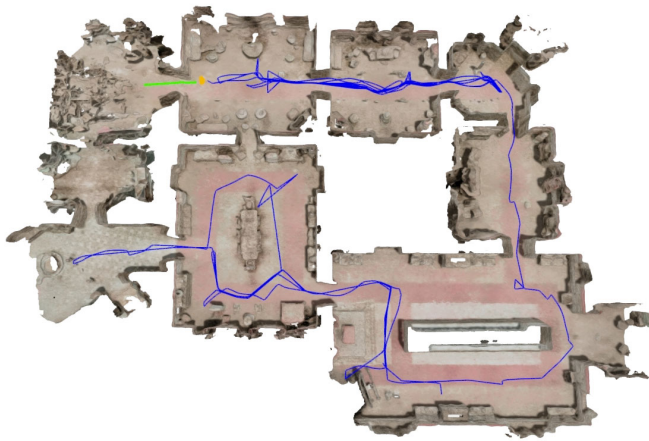
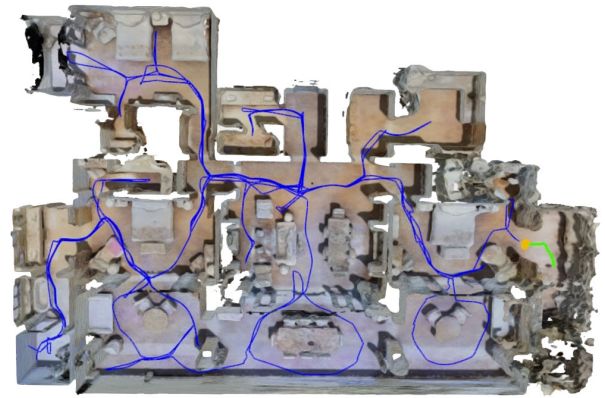


Fig. 5: The complex scene with severe occlusions addresses challenges to trade-offs between efficient exploration and accurate reconstruction.

<sup>2</sup>the percentage of points whose nearest distance is within 5cm

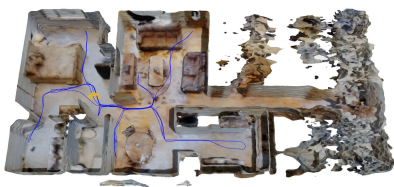


(a) MP3D-Z6MFQ (22 rooms with 90.97% of completeness)

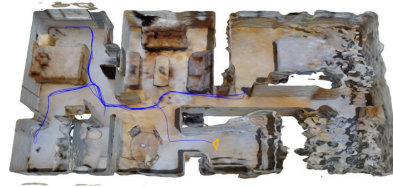


(b) MP3D-q9vSo (22 rooms with 92.93% of completeness)

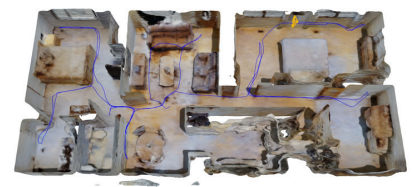
Fig. 6: The active reconstruction results of large-scale scenes of the Matterport3D dataset.



(a) Random Voronoi vertex (64.78%)



(b) Best anchored vertex (73.93%)



(c) Ours (91.46%)

Fig. 7: The reconstruction results using different strategies of the next-best-view selection. (Gibson-Cantwell in 2000 steps)

complicated environment (MP3D-YmJkq) with severe occlusions and multiple narrow navigable pathways is still challenging. The exploration completeness is the worst among all tested scenes even though it still beats the competitors.

### C. Reconstructing large-scale scenes

Besides the test split in [13], we further validate the efficacy of the proposed method on the largest three scenes in test/val splits of Matterport3D dataset. Each scene consists of over 20 rooms. As illustrated in Fig. 6 and Fig. 1, the proposed method can achieve over 90% completeness (including ceiling) and reconstruct fine-grained surface meshes in at most 10,000 steps. The integration of the topology map and the neural map assures thorough exploration and accurate reconstruction.

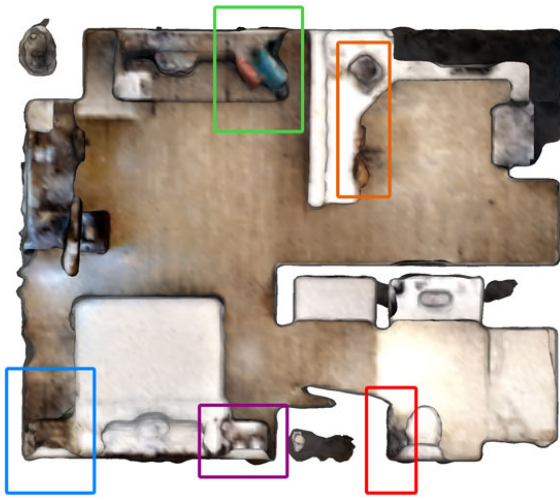
### D. Ablation studies

We conduct ablation studies to study the impact of different modules given the proposed strategies.

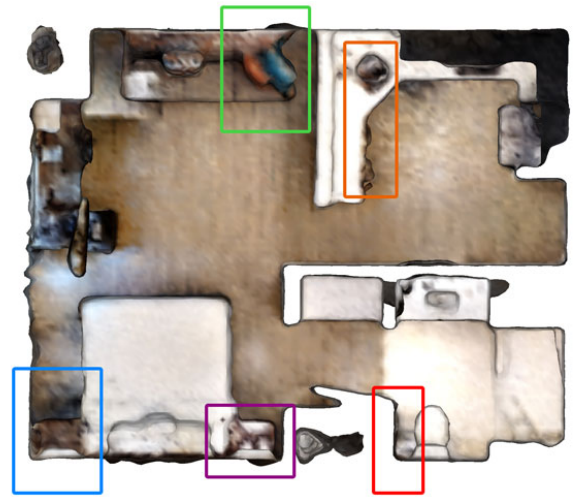
*a) Voronoi vertex selection:* The proposed method takes account of both the accessible regions of interest and the visible regions of interest by anchoring the uncertain areas to the Voronoi vertices. We compare different strategies for the next-best-view selection. As illustrated in Fig. 7, randomly selecting the Voronoi vertex and filtering out previously visited ones lead to 64.78% of completeness, while

chasing the Voronoi vertex anchored by the most uncertain sub-areas leads to 73.93% of completeness. The final result of ours prioritizes the anchored Voronoi vertex within the local horizon and follows a near-to-far order to traverse the vertices that have not been visited. The hierarchical framework leads to 91.46% of completeness.

*b) Visible regions of interest:* The proposed method targets an active neural mapping setting beyond pursuing the overall coverage of the environment. Therefore, we design delicate strategies concerning both local-global hierarchies and visible-accessible balance so the system can reconstruct accurate scenes with efficient traversal. As illustrated in Fig. 8, we validate the efficacy of our visibility guidance from the uncertainty quantification method. The closeups validate that rotating towards the uncertain areas once approaching the Voronoi vertex leads to better details of geometry and appearance. The proposed method is designed to adopt a meticulous strategy near the target Voronoi vertex to search for unseen areas. This is due to the fact that the Voronoi vertex is the informative intersection of multiple pathways, where careful decisions should be made. The updated Voronoi graph on the other side defines the accessible regions of interest that guides the agent towards unvisited places while guaranteeing safe path planning during exploration. The collaboration of both sides achieves a nice trade-off between exploration efficiency and reconstruction



(a) Without visibility guidance



(b) With visibility guidance

Fig. 8: Visibility guidance leads to the fine-grained reconstruction of details compared to accessibility-only exploration.

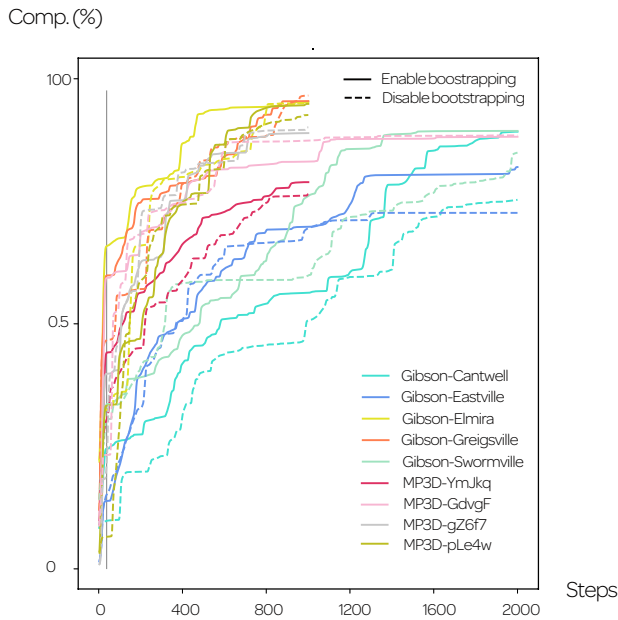


Fig. 9: The bootstrap strategy (36 steps as the grey line for a complete rotation) leads to faster coverage at the beginning given a more confident initial state.

accuracy.

*c) Bootstrap strategy:* In the very beginning, we start by executing a complete rotation before the agent starts to explore. This strategy ensures a more confident understanding of the surrounding environment by sacrificing 36 steps. As is shown in Fig. 9, this simple strategy leads to faster ascendance of coverage in small scenes or when the agent is initialized at a complicated crossing. Starting with better confidence leads to consistently better completeness in most cases. The final completeness with or without bootstrap

shows a similar ratio for all small scenes as the topology map enforces thorough exploration.

## VI. CONCLUSION

In this paper, we present a NeRF-based active mapping method. Taking advantage of the topology within the neural map and a hybrid network architecture, the proposed method greatly enhances the scalability of the active neural mapping problem. Taking visible and accessible regions of interest into consideration in a hierarchical framework, the proposed method achieves a nice trade-off between exploration efficiency and reconstruction accuracy with promising real-time capability. The experiments validate the state-of-the-art performance. Future potentials include the integration of more structural information about the environments such as semantics and relations between objects that better define the local similarity for efficient and accurate exploration and reconstruction.

## VII. ACKNOWLEDGEMENTS

The work is supported by NSFC (U22A2061, 62176010) and 230601GP0004.

## REFERENCES

- [1] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020, pp. 7462–7473.
- [2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conf. on Computer Vision (ECCV)*, 2020, pp. 405–421.
- [3] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graphics*, vol. 41, no. 4, pp. 1–15, 2022.
- [4] G. Yüce, G. Ortiz-Jiménez, B. Besbinar, and P. Frossard, "A structured dictionary perspective on implicit neural representations," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19 228–19 238.

- [5] D. B. Lindell, D. Van Veen, J. J. Park, and G. Wetzstein, "Bacon: Band-limited coordinate networks for multiscale scene representation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16252–16262.
- [6] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-nerf: Anti-aliased grid-based neural radiance fields," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 19697–19705.
- [7] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15838–15847.
- [8] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 19729–19739.
- [9] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *Intl. Conf. on 3D Vision (3DV)*. IEEE, 2017, pp. 667–676.
- [10] C. Connolly, "The determination of next best views," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, vol. 2. IEEE, 1985, pp. 432–435.
- [11] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *IEEE Intl. Sym. on Computational Intelligence in Robotics and Automation (CIRA)*, 1997, pp. 146–151.
- [12] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon "next-best-view" planner for 3d exploration," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1462–1468.
- [13] Z. Yan, H. Yang, and H. Zha, "Active neural mapping," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 10981–10992.
- [14] H. Wang, J. Wang, and L. Agapito, "Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13293–13302.
- [15] E. Sandström, Y. Li, L. Van Gool, and M. R. Oswald, "Point-slam: Dense neural point cloud-based slam," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 18433–18444.
- [16] L. Goli, C. Reading, S. Sellán, A. Jacobson, and A. Tagliasacchi, "Bayes' rays: Uncertainty quantification for neural radiance fields," *arXiv preprint arXiv:2309.03185*, 2023.
- [17] S. Dong, K. Xu, Q. Zhou, A. Tagliasacchi, S. Xin, M. Niebner, and B. Chen, "Multi-robot collaborative dense scene reconstruction," *ACM Trans. Graphics*, vol. 38, no. 4, pp. 1–16, 2019.
- [18] K. Ye, S. Dong, Q. Fan, H. Wang, L. Yi, F. Xia, J. Wang, and B. Chen, "Multi-robot active mapping via neural bipartite graph matching," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14839–14848.
- [19] S. Shen, N. Michael, and V. Kumar, "Autonomous indoor 3d exploration with a micro-aerial vehicle," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2012, pp. 9–15.
- [20] B. Zhou, Y. Zhang, X. Chen, and S. Shen, "Fuel: Fast uav exploration using incremental frontier structure and hierarchical planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 779–786, 2021.
- [21] L. Schmid, M. Pantic, R. Khanna, L. Ott, R. Siegwart, and J. Nieto, "An efficient sampling-based method for online informative path planning in unknown environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1500–1507, 2020.
- [22] T. Dang, C. Papachristos, and K. Alexis, "Visual saliency-aware receding horizon autonomous exploration with application to aerial robotics," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2526–2533.
- [23] M. Selin, M. Tiger, D. Duberg, F. Heintz, and P. Jensfelt, "Efficient autonomous exploration planning of large-scale 3d environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1699–1706, 2019.
- [24] A. Batinovic, A. Ivanovic, T. Petrovic, and S. Bogdan, "A shadowcasting-based next-best-view planner for autonomous 3d exploration," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2969–2976, 2022.
- [25] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," in *Intl. Conf. on Learning Representations (ICLR)*, 2019.
- [26] T. Gervet, S. Chintala, D. Batra, J. Malik, and D. S. Chaplot, "Navigating to objects in the real world," *Science Robotics*, vol. 8, no. 79, p. eadf6991, 2023.
- [27] T. Chen, S. Gupta, and A. Gupta, "Learning exploration policies for navigation," in *Intl. Conf. on Learning Representations (ICLR)*, 2019.
- [28] A. Sridhar, D. Shah, C. Glossop, and S. Levine, "Nomad: Goal masked diffusion policies for navigation and exploration," in *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.
- [29] Z. A.-H. Santhosh Kumar Ramakrishnan and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *European Conf. on Computer Vision (ECCV)*, 2020.
- [30] Z. Yan, Y. Tian, X. Shi, P. Guo, P. Wang, and H. Zha, "Continual neural mapping: Learning an implicit scene representation from sequential observations," in *Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 15782–15792.
- [31] E. Sucar, S. Liu, J. Ortiz, and A. J. Davison, "imap: Implicit mapping and positioning in real-time," in *Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 6229–6238.
- [32] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "Nice-slam: Neural implicit scalable encoding for slam," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12786–12796.
- [33] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, "Go-slam: Global optimization for consistent 3d instant reconstruction," in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 3727–3737.
- [34] X. Pan, Z. Lai, S. Song, and G. Huang, "Activenerf: Learning where to see with uncertainty estimation," in *European Conf. on Computer Vision (ECCV)*. Springer, 2022, pp. 230–246.
- [35] Y. Ran, J. Zeng, S. He, J. Chen, L. Li, Y. Chen, G. Lee, and Q. Ye, "Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1125–1132, 2023.
- [36] Z. Feng, H. Zhan, Z. Chen, Q. Yan, X. Xu, C. Cai, B. Li, Q. Zhu, and Y. Xu, "Naruto: Neural active reconstruction from uncertain target observations," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 21572–21583.
- [37] H. Choset and K. Nagatani, "Topological simultaneous localization and mapping (slam): toward exact localization without explicit localization," *IEEE Trans. Robotics and Automation*, vol. 17, no. 2, pp. 125–137, 2001.
- [38] E. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [39] J. Kim, F. Zhang, and M. Egerstedt, "A provably complete exploration strategy by constructing voronoi diagrams," *Autonomous Robots*, vol. 29, pp. 367–380, 2010.
- [40] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys (CSUR)*, vol. 23, no. 3, pp. 345–405, 1991.
- [41] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *Intl. J. of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.
- [42] T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák, "Neural importance sampling," *ACM Trans. Graphics*, vol. 38, no. 5, pp. 1–19, 2019.
- [43] G. Georgakis, B. Bucher, A. Arapin, K. Schmeckpeper, N. Matni, and K. Daniilidis, "Uncertainty-driven planner for exploration and navigation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022.
- [44] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A platform for embodied AI research," in *Intl. Conf. on Computer Vision (ICCV)*, 2019.
- [45] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9068–9079.