Beyond Frontiers: Scene-Anomaly Guided Autonomous Exploration

Akash Kumbar[1] Abhinav Raundhal[1] Madhava Krishna[1]

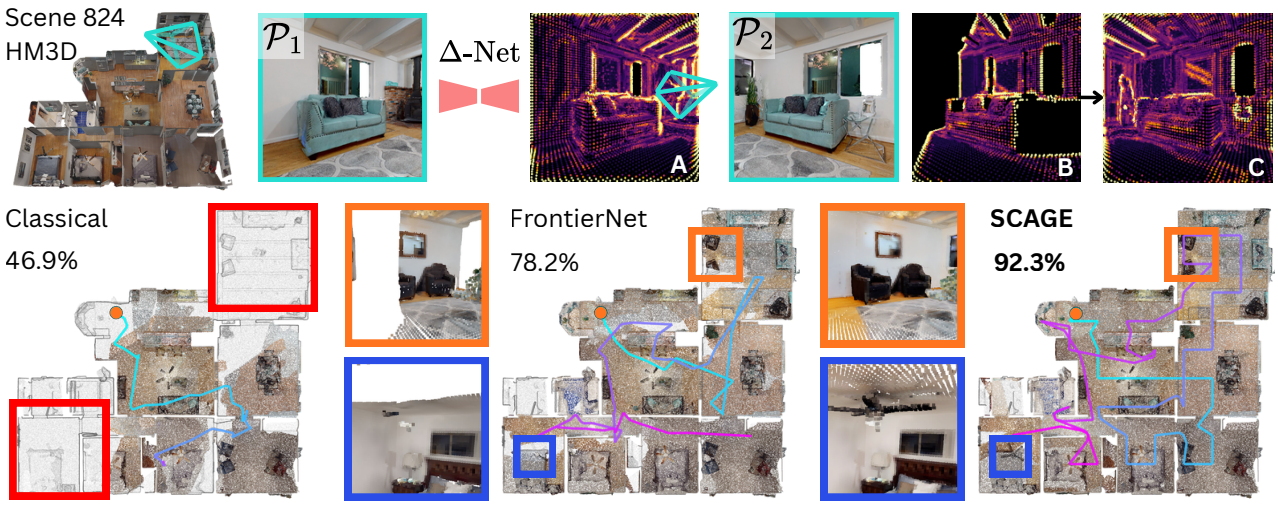

Fig. 1: **Top.** Our anomaly-guided exploration. Given a 3D point cloud at pose $\mathcal{P}_1$ (RGB images shown for context), our network Δ-Net predicts *scene anomalies* (A) (shown in yellow on the point cloud), representing unexplored or poorly reconstructed areas. This anomaly signal guides the robot to an optimal next-best-view ($\mathcal{P}_2$). The anomalies as seen from $\mathcal{P}_2$, prior to obtaining the new point cloud, are shown in (B), which get resolved after obtaining the point cloud from this view as shown in (C). **Bottom.** Trajectory and mapping comparison on an HM3D scene from a starting position marked by an orange dot. Classical methods leave large sections unexplored (red boxes), and FrontierNet overlooks local geometric details. In contrast, our method, SCAGE achieves the highest volumetric coverage (92.3%) while simultaneously producing superior, high-fidelity 3D reconstructions (orange and blue insets highlight the reconstruction quality of fan and chair which FrontierNet fails to achieve).

***Abstract*— Autonomous exploration of unknown 3D environments is traditionally driven by coverage-maximizing geometric heuristics. However, these methods typically determine exploration targets without considering the underlying structural context. This leads to inefficient trajectories often limiting the fidelity of the final 3D reconstruction. To bridge the gap between spatial coverage and reconstruction quality, we introduce a novel paradigm: reframing exploration as a geometric anomaly minimization problem. We present SCAGE: SCene Anomaly Guided Exploration, a novel autonomous exploration framework that operates directly on unstructured 3D point clouds. Instead of blindly chasing volumetric boundaries, we equip the robot with a foundational understanding of standard indoor architecture. As the robot navigates, it continuously evaluates its live 3D observations against these learned expectations. When the incoming geometry contradicts the learned priors of a typical indoor environment, such as a fragmented wall or a partial table, the system flags these regions as *scene anomalies*. These geometric inconsistencies act as a guiding signal, naturally drawing the robot to investigate and resolve these structural anomalies from optimal vantage points. By actively targeting poorly reconstructed regions rather than just empty space, our approach seamlessly couples spatial discovery with high-fidelity mapping. Extensive evaluations demonstrate that SCAGE achieves superior volumetric coverage (∼90% in all scenes) and higher 3D reconstruction quality compared to state-of-the-art baselines.**

[1]Robotics Research Center, IIIT-Hyderabad, India
†Project page: https://beyondfrontiers.github.io

I. INTRODUCTION

Autonomous exploration requires a mobile robot to navigate through unknown 3D environments to build digital maps, search for objects, or gather critical information. This capability is fundamental to a wide range of real-world applications, including search and rescue [1], agricultural monitoring [2], and infrastructure inspection [3]. Efficient autonomous exploration, whether aimed at maximizing mapped volume, enriching semantic understanding, or improving 3D reconstruction quality, ultimately boils down to determining the optimal sequence of poses.

Historically, exploration strategies have relied on geometric heuristics, either by driving the robot toward the boundaries between mapped and unmapped space [4], [5], or by evaluating candidate poses against occupancy-derived metrics and selecting the most informative ones [6], [7]. In both cases, utility is computed by ray-casting into the occupancy grid and therefore reflects only the volume of unknown space a viewpoint reveals. Since unknown regions are undifferentiated and observed surfaces are considered resolved once marked occupied, these methods reward large open voids over occluded or sparsely sampled geometry, yielding trajectories that expand coverage while leaving structurally significant regions poorly reconstructed.

Recent learning-based methods attempt to overcome these

limitations by using neural networks to predict the potential value of unmapped regions. While effective, these approaches typically rely on 2D visual cues [8], [9] or external foundational models [10]. More importantly, they maintain a narrow focus on maximizing spatial coverage. This inadvertently decouples the act of exploring a space from the ultimate goal of generating a high-fidelity 3D reconstruction of it, leading to maps that may be complete in volume but lacking in geometric detail or structural integrity (as shown in Fig. 1). While recent research has begun utilizing scene completion networks [11] and semantic priors [12] to predict information gain through extrapolating unobserved volumes, these explicit hallucination methods often act as weak exploration indicators due to their limited capability to reliably complete highly complex indoor scenes.

In this paper, we introduce a novel paradigm in autonomous exploration: rather than simply searching for volumetric boundaries, we reframe exploration as a geometric anomaly-minimization problem. We ask: *Which part of our current reconstruction fails to meet our learned architectural priors?* By targeting these geometric contradictions, we naturally guide the robot toward unresolved or occluded regions that are otherwise neglected.

Typical indoor environments exhibit predictable geometric distributions like flat walls, orthogonal intersections, and standard furniture layouts. Incomplete, sparse, or occluded observations inherently deviate from these distributions creating what we define as *scene anomalies*, localized regions where the observed 3D geometry contradicts established architectural logic. We propose SCAGE: SCene Anomaly Guided Exploration, a novel exploration framework that operates directly on 3D point clouds to identify these anomalies. At the core of our framework is Δ-Net, a denoising autoencoder based on the Point Transformer V3 [13] architecture. Designed to learn the geometric priors of typical indoor structures, Δ-Net explicitly predicts the required point-wise distribution shift (the Δ) between a partial observation and its expected architectural form. Crucially, we train the network on localized point cloud chunks rather than entire spatial layouts, to focus on learning the fundamental structural features that compose typical indoor environments.

During online exploration, the network predicts a point-wise geometric deviation vector for incoming observations. Highly anomalous regions are clustered to compute adaptive 6-DoF observation viewpoints dynamically. Guided by a two-step lookahead utility planner, this anomaly-driven approach naturally steers the robot toward unexplored and under-reconstructed areas, allowing the system to maximize coverage and minimize reconstruction errors jointly.

In summary, the key contributions of our work are:

- We reframe 3D autonomous exploration as an anomaly minimization problem, introducing a novel framework, SCAGE, that uses geometric deviation to drive navigation.
- We design a denoising autoencoder Δ-Net that efficiently learns the structural priors of standard indoor environments from localized point cloud chunks.
- We demonstrate through extensive evaluations that SCAGE achieves superior volumetric coverage ($\sim$15% improvement) and higher 3D reconstruction quality (half the error) as compared to state-of-the-art exploration baselines.

II. RELATED WORK

Geometry-Based Exploration. Traditional exploration relies on geometric heuristics to identify optimal viewpoints. Frontier-based methods extract boundaries between known and unmapped regions to direct navigation [4], [5]. Alternatively, sampling-based approaches evaluate candidate poses against metrics like map entropy and uncertainty to maximize volumetric coverage [6], [7]. Hybrid approaches bridge these paradigms, combining the targeted efficiency of frontiers with the evaluative power of sampling. This is achieved either by evaluating frontier-sampled viewpoints via entropy and travel-time utility functions [14], or by pairing global frontier exploration with local Next-Best-View Planning (NBVP) [15]. Recent hierarchical frameworks further optimize these strategies for large-scale 3D environments [16]. Although these methods maximize mapped volume, they overlook the underlying structural context of the environment often causing myopic over-exploration of local regions at the expense of discovering new spaces.

Learning-Based Exploration. To overcome the limitations of purely geometric methods, recent literature increasingly leverages learning-based vision algorithms. For instance, several approaches estimate the information gain of candidate frontiers using 3D occupancy prediction models or scene completion networks [11], [12], while others bypass dense 3D maps entirely by predicting frontier utility directly from 2D visual cues [8]. Emerging frameworks explore novel map structures, utilizing neural implicit representations [17], [18] and 3D Gaussians [19], [20], [21]. Beyond pure geometry, incorporating appearance data such as object-level semantics into 3D representations has proven effective for refining trajectory and viewpoint evaluation [22], [23], [24]. Finally, alternative paradigms treat exploration as a decision-making problem, employing reinforcement learning on color images [10], [25], [26], or utilizing vision foundation models and LLMs to guide human-like navigation [27], [28], [29].

Redefining Frontiers for Exploration. Unlike existing learning-based methods that rely on 2D visual cues or voxel-grid completions which focus on maximizing coverage, our approach operates directly on 3D point clouds. We shift the objective from simply finding empty space to detecting structural anomalies, guiding the robot toward areas that deviate from the expected geometry of typical indoor environments. This allows our system to jointly maximize coverage and minimize reconstruction errors by naturally targeting unresolved or occluded regions that standard volume-driven methods ignore.

III. SCAGE OVERVIEW

We present an overview of the proposed SCAGE pipeline in Fig. 2. The framework consists of two primary phases.

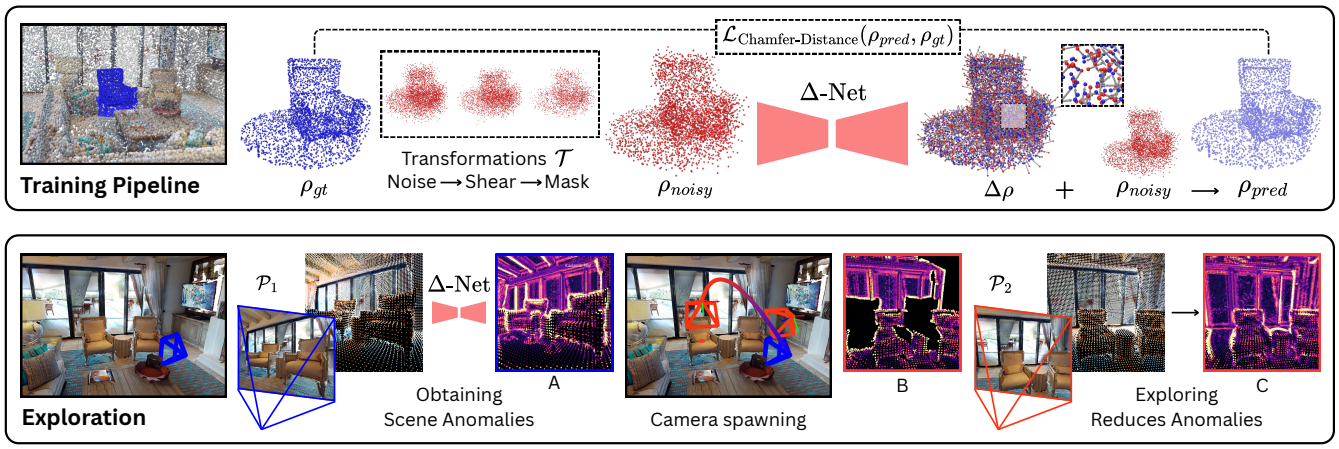

Fig. 2: **Overview of SCAGE. Top.** The Training Pipeline. Given point clouds of indoor scenes, we partition them to obtain ground truth chunks ρ_{gt}. We apply transformations $\mathcal{T}$ (e.g., Noise, Shearing, Masking) to simulate partial observability, yielding ρ_{noisy}. Our Δ-Net architecture processes this corrupted input to predict a point-wise distribution shift $\Delta\rho$, visualized as per-point arrows mapping the noisy input ρ_{noisy} toward the ground truth ρ_{gt}. The predicted $\Delta\rho$ is added to ρ_{noisy} to obtain ρ_{pred}. The network is trained by minimizing the $\mathcal{L}_{\text{Chamfer-Distance}}$ between ρ_{pred} and ρ_{gt}. **Bottom.** The Exploration phase. Given a partial 3D observation from an initial pose $\mathcal{P}_1$, Δ-Net predicts the distribution shift, which is thresholded to identify *Scene Anomalies* (A), with yellow representing high-anomaly regions. Cameras are spawned to target these anomalies, generating candidate 6-DoF viewpoints (red frustums). The robot then navigates to the optimal pose $\mathcal{P}_2$. The anomalies, as seen from $\mathcal{P}_2$ (B), are resolved (C) once the new point cloud is captured, enabling high-fidelity 3D reconstruction.

First, in the training pipeline (Section III-A), we train a point-based denoising autoencoder to learn the underlying structural priors of typical indoor environments. Second, during the exploration phase (Section III-B), we leverage the point-wise distribution shift from these learned priors to identify unseen regions. The magnitude of these structural deviations serves as an anomaly score to drive autonomous exploration.

A. Learning what indoor scenes look like

Data Preparation. Our objective is to capture the geometric distribution of complete, fully observed indoor scenes. To achieve this, we partition 3D point clouds of various indoor scenes from the Matterport3d dataset [30] into localized chunks $\rho_{gt} \in \mathbb{R}^{N\times 3}$ ($N = 4096$). To construct supervised training pairs that teach the network to map degraded observations back to their idealized structural form, we apply a series of stochastic transformations $\mathcal{T}$ to these ground-truth chunks. Specifically, we corrupt the input by injecting Gaussian noise with varying standard deviations, applying anisotropic scaling, and utilizing random point masking (as illustrated in fig. 2) to simulate geometric distortions, partial coverage, and occlusions that arise from real-world depth sensing, yielding noisy input chunks: $\rho_{noisy} = \mathcal{T}(\rho_{gt})$.

Model Architecture and Training. Our architecture is built upon the Point Transformer V3 [13] backbone, which efficiently processes 3D data, serializing unstructured 3D point sets into 1D sequences via space-filling curves and applying localized serialized attention. Given a noisy input point cloud chunk $\rho_{noisy} \in \mathbb{R}^{N\times 3}$, the proposed network, Δ-Net employs a hierarchical encoder-decoder structure, which we train from scratch.

The encoder progressively downsamples the point resolution into a compact 512-channel latent representation. Crucially, we enforce a strict information bottleneck by excluding any skip connections between the encoder and decoder, preventing the network from bypassing this compressed representation and directly copying input features. This design choice forces the autoencoder to internalize the geometric priors of typical indoor structures rather than overfit to the specific corruptions present in the training data. The symmetric decoder then upsamples this latent representation back to the original point resolution, and a final Multi-Layer Perceptron (MLP) head predicts a point-wise distribution shift vector $\Delta\rho \in \mathbb{R}^{N\times 3}$. The final denoised coordinates are then computed as $\rho_{pred} = \rho_{noisy} + \Delta\rho$.

During training, Δ-Net is optimized to reconstruct the underlying geometry by minimizing the Chamfer Distance between the predicted chunk ρ_{pred} and the ground truth chunk ρ_{gt}.

$$\mathcal{L}_{CD}(\rho_{pred}, \rho_{gt}) = \sum_{x\in\rho_{pred}} \min_{y\in\rho_{gt}} \|x-y\|_2^2 + \sum_{y\in\rho_{gt}} \min_{x\in\rho_{pred}} \|x-y\|_2^2 \tag{1}$$

This bidirectional loss penalizes spatial divergence, forcing the autoencoder to map anomalous points back to the learned distribution of standard indoor architecture. The resulting point-wise distribution shift vector $\Delta\rho$ captures how strongly each observed region deviates from the learned geometric prior, and is leveraged during inference to guide the robot toward structurally incomplete areas, as described next.

B. Exploration using Scene Anomalies

Detecting Scene Anomalies. During exploration, starting from the initial camera pose, we obtain a depth map and backproject it into 3D to obtain a point cloud of the observed scene. In subsequent iterations, the newly backprojected points are merged with the accumulated scene representation to form a global point set P.

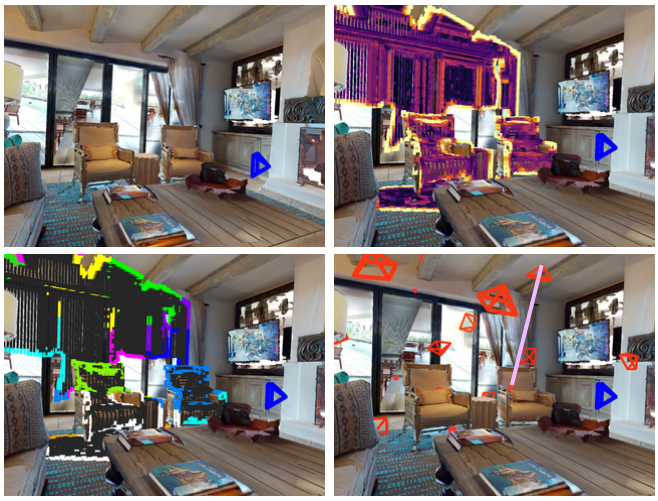

Fig. 3: **Viewpoint Generation from Scene Anomalies**. Top Left: A part of a scene with a camera shown with a blue frustum. Top Right: Per-point anomaly scores obtained from this camera view, with high-anomaly regions shown in yellow. Bottom Left: Anomalies are clustered into regions shown in different colors. Bottom Right: For each cluster, the cluster anchor is shown with a red dot and the corresponding candidate 6-DoF viewpoints (red frustums) are spawned facing each anomaly cluster. They are placed at a standoff distance from the cluster anchor (shown with a pink line for one of the cameras), adapting to the spatial extent of the cluster. Larger clusters (e.g., the wall boundary) produce viewpoints at greater standoff distances, while compact clusters are observed from nearby, ensuring each anomalous region is fully captured within the camera's field of view.

To maintain uniform spatial density and control computational cost, we apply voxel downsampling with resolution 0.05m to obtain a filtered point cloud $\tilde{P}$. The downsampled point cloud is then partitioned into overlapping localized chunks, each processed independently by Δ-Net.

For each chunk $\rho \in \mathbb{R}^{N\times 3}$, the network predicts a point-wise distribution shift $\Delta\rho \in \mathbb{R}^{N\times 3}$. For a point $p_i \in \rho$, let its predicted deviation vector be Δp_i. If a point belongs to multiple overlapping chunks, we aggregate its deviation magnitude by taking the maximum response. This produces a global deviation field over $\tilde{P}$.

We then compute the globally normalized anomaly score

$$s_i = \frac{\|\Delta p_i\|_2 - \min_j \|\Delta p_j\|_2}{\max_j \|\Delta p_j\|_2 - \min_j \|\Delta p_j\|_2}, \quad s_i \in [0,1] \tag{2}$$

where the extrema are evaluated over all points in $\tilde{P}$.

These per-point anomaly scores quantify the degree to which local geometry deviates from the learned structural prior and serve as the signal for exploration. To reduce redundant computation, we skip re-processing points lying in low-anomaly neighborhoods, thereby concentrating inference on structurally uncertain and unexplored regions.

Explore the unseen. To convert these anomaly signals into navigable targets, we extract the anomalous set $\mathcal{A} = \{p_i \in \tilde{P} \mid s_i > \tau\}$, where $\tau = 0.5$ is an empirically selected threshold. Because these anomalous points often represent contiguous unobserved regions or boundaries, we group $\mathcal{A}$ into distinct target clusters C_k, using ball-query based clustering.

Viewpoint Generation. For each cluster C_k, we compute a 6-DoF observation viewpoint that directly addresses the detected anomaly (as illustrated in Fig. 3). We first obtain the geometric center (anchor) a_k and extract the local surface normal $\hat{n}_k$ toward the sensor origin to ensure the resulting viewpoint lies in observed free space.

The observation distance must adapt to the spatial extent of each anomaly: small clusters can be observed from nearby, while spatially extended regions require the camera to retreat so the full cluster remains visible. Concretely, we compute d_{standoff}, the minimum distance at which a cluster of spatial radius $r_k = max_{p_j \in C_k} \|p_j - a_k\|_2$ subtends an angle no greater than the camera's horizontal field of view:

$$d_{\text{standoff}} = \max\left(d_{\min}, \frac{r_k}{\tan(\theta_{\text{FOV}}/2)}\right) \tag{3}$$

where d_{min} enforces a physical safety margin. The candidate camera position is then $\mathcal{P}_{cam,k} = a_k + d_{\text{standoff}} \cdot \hat{n}_k$, with the camera yaw oriented toward a_k, fully defining a preliminary 6-DoF target pose T_k.

Target Refinement and Selection. Because dense anomaly clustering can produce redundant or kinematically unreachable targets, all candidates undergo a rigorous refinement phase. Targets are spatially consolidated to merge highly overlapping views, followed by free-space and line-of-sight validation against a volumetric occupancy grid. Then, for each surviving candidate T_k, we evaluate its exploratory value using the following utility function:

$$u(T_k) = \frac{g(T_k) \cdot \bar{s}_k}{\max(\|\mathcal{P}_{\text{robot}} - \mathcal{P}_{\text{cam},k}\|_2, \epsilon)} \tag{4}$$

where $g(T_k)$ is the expected volumetric information gain computed via raycasting, $\bar{s}_k$ is the mean anomaly score of cluster C_k, and the denominator represents the Euclidean traversal cost from the robot's current position. This formulation balances three competing objectives: preferring views that reveal new volume, that resolve high-anomaly regions, and that minimize travel distance. We ablate this design in Table II, demonstrating all three terms are necessary for strong performance. The robot navigates to the utility-maximizing target via a collision-free target path planned using the RRT* algorithm [31].

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Implementation Details. In order to ensure that Δ-Net learns geometric priors from structurally complete environments, we curate a training set of 75 high-fidelity scenes from the Matterport3D dataset. These scenes were specifically selected for their high structural integrity, actively filtering out scans with significant artifacts or missing geometry. By training Δ-Net exclusively on these clean, continuous manifolds from Matterport3D, and evaluating on unseen HM3D validation scenes, we demonstrate the robust cross-dataset generalization of our learned prior. The network was trained for 600 epochs on a single NVIDIA RTX A6000 GPU.

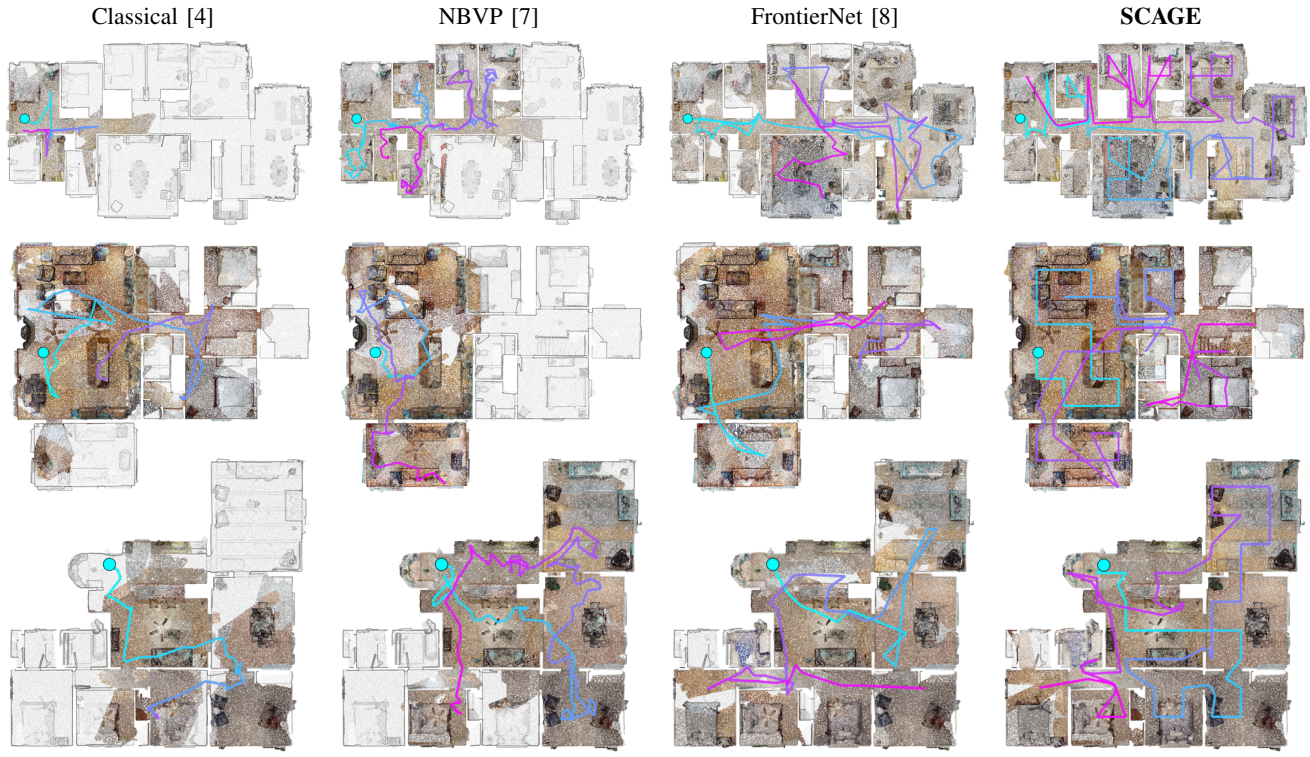

Fig. 4: **Qualitative Comparison. (Coverage and Trajectory).** Exploration results of our method compared against baselines on 3 scenes (top to bottom: 804, 812, 824). Starting location is marked with a cyan point, unexplored regions are shown in grey. The trajectory is shown in a gradient from cyan at the start to pink towards the end. Our method, SCAGE, achieves the maximum coverage across all baselines on all scenes. Though FrontierNet [8] seems to achieve high coverage, it often just visits the room at its entrance, failing to capture structural elements that compose the room like chairs or beds. Whereas, by modeling the exploration as an anomaly minimization problem, we achieve both maximum coverage and good quality reconstruction, capturing all the fine details that construct indoor scenes.

Benchmark Standardization and Simulation. As recently highlighted by FrontierNet [8], the evaluation of autonomous exploration systems is frequently hindered by the lack of standardized test protocols and a reliance on limited, disparate scenarios. Concurring with this assessment, we adopt their proposed evaluation framework to ensure a transparent comparison. We evaluate our approach using the identical subset of HM3D validation scenes, spanning diverse geometric complexities and spatial scales, and enforce the same maximum number of simulation steps per episode. By freezing the validation split, episode horizons, and sensor specifications, we isolate our learned priors as the sole source of performance gains. We simulate camera viewpoints and render images using Open3D, generating depth maps at a resolution of 480×480. The sensor is configured with a field of view (FOV) of $77.32° \times 77.32°$ and a maximum depth range of 5m. Depth input is processed in two variants: (1) perfect depth rendered directly from the simulation and (2) noisy depth predicted by Metric3D v2 [32]. We maintain a volumetric occupancy map using Octomap and rely on the Open Motion Planning Library (OMPL) for low-level 3D path generation.

B. Results

Baselines. We benchmark the exploration efficiency of SCAGE against three established frameworks spanning both traditional and learning-based paradigms: our implementation of classic frontier-based method [4], the sampling-based Next-Best-View Planner (NBVP) [7], and the recent state-of-the-art visual exploration approach, FrontierNet [8]. To guarantee a fair and rigorous comparison, we utilize the official baseline implementations and strictly adhere to the evaluation configurations defined by [8], specifically testing across their identical subset of 10 validation meshes and enforcing their exact scene-specific maximum step budgets. We evaluate the baselines with perfect depth from the simulator and our method with both depth from simulator and Metric3D v2 [32].

Quantitative Evaluation We quantify exploration performance using two primary metrics: Volumetric Coverage and Reconstruction Quality (measured via Chamfer Distance) averaged over multiple runs from several starting positions. As summarized in Table I, SCAGE consistently outperforms all baselines, achieving a maximum volumetric coverage of approximately 90% across all evaluated scenes, a $\sim$15% improvement over the best baseline FrontierNet [8]. This demonstrates our method's capacity to thoroughly navigate and map complex layouts without prematurely terminating or leaving significant regions unexplored.

Crucially, this expansive coverage does not come at the expense of geometric detail. Our approach achieves the best overall reconstruction quality, as evidenced by yielding the lowest Chamfer Distance among all evaluated methods. As compared to the best baseline we achieve less than half the

	Method	Scene ID 804	807	812	824	827	834	854	876	879	880	Mean
Volume Coverage ↑	Classic [4]	30.9 ± 5.3	45.6 ± 2.2	64.5 ± 4.8	48.9 ± 2.3	59.1 ± 7.9	49.7 ± 5.3	51.3 ± 4.4	44.2 ± 8.8	50.2 ± 4.1	62.4 ± 5.3	50.7
	NBVP [7]	37.1 ± 6.7	50.1 ± 3.2	83.7 ± 4.1	64.3 ± 4.8	76.7 ± 4.1	64.0 ± 9.4	80.3 ± 2.3	61.4 ± 10.3	63.1 ± 8.4	47.2 ± 3.1	62.8
	FrontierNet [8]	55.4 ± 7.1	56.1 ± 6.1	81.8 ± 2.2	72.4 ± 6.2	73.4 ± 9.1	74.5 ± 10.3	93.1 ± 5.3	75.9 ± 8.5	70.8 ± 10.1	69.2 ± 10.1	72.3
	SCAGE	92.5 ± 1.0	84.5 ± 0.3	88.8 ± 0.5	93.0 ± 0.5	90.0 ± 0.5	84.1 ± 0.5	91.4 ± 3.5	91.0 ± 0.1	90.4 ± 0.1	91.6 ± 0.8	89.7
	SCAGE*	90.1 ± 1.9	81.5 ± 2.4	85.8 ± 1.8	90.0 ± 2.1	87.3 ± 3.5	83.1 ± 1.1	92.1 ± 4.5	92.9 ± 1.6	88.4 ± 0.3	90.1 ± 3.8	88.1
Chamfer($\times 10^{-3}$) ↓	Classic [4]	12.0 ± 4.0	11.0 ± 3.5	3.5 ± 1.0	5.5 ± 2.0	6.0 ± 2.5	4.5 ± 2.0	1.5 ± 1.0	4.5 ± 2.0	6.5 ± 2.5	10.5 ± 4.0	6.55
	NBVP [7]	9.5 ± 3.5	8.5 ± 3.0	1.8 ± 0.8	3.5 ± 1.5	3.8 ± 1.8	2.8 ± 1.5	0.8 ± 0.5	2.8 ± 1.2	4.5 ± 2.0	8.0 ± 3.5	4.60
	FrontierNet [8]	7.1 ± 3.4	6.5 ± 2.9	0.9 ± 0.3	1.9 ± 1.4	2.1 ± 1.5	1.4 ± 0.9	0.6 ± 0.2	1.5 ± 1.2	2.9 ± 1.8	5.3 ± 4.3	3.02
	SCAGE	1.2 ± 0.8	0.5 ± 0.3	0.7 ± 0.2	1.0 ± 0.6	0.6 ± 0.4	1.6 ± 0.4	0.5 ± 0.0	0.6 ± 0.1	0.6 ± 0.1	1.7 ± 0.5	0.91
	SCAGE*	2.4 ± 1.2	1.2 ± 0.6	1.6 ± 0.5	2.1 ± 1.0	1.3 ± 0.7	3.1 ± 0.9	1.3 ± 0.3	1.4 ± 0.4	1.3 ± 0.3	3.2 ± 1.1	1.89

TABLE I: **Quantitative Comparison.** Results comparing Volumetric Coverage and Reconstruction Quality quantified by Chamfer Distance. Our method achieves the maximum coverage (∼90%, a substantial ∼15% improvement on an average over the best baseline) in all scenes, showing its capability to explore without missing any parts of the scene. More importantly, along with the coverage, we achieve the best reconstruction quality, as shown by lowest chamfer distance. We obtain less than half the reconstruction error, on an average, as compared with the best baseline. Notice the low standard deviation of our method, highlighting its robustness towards different initial positions across multiple runs. The 3 digit numbers in the first row are scene IDs. Results in the first 4 rows are using depth from the simulator including SCAGE, whereas the last row SCAGE* uses the depth estimated by Metric3D v2 [32]. Red, orange, and yellow highlights indicate the first, second, and third best performing technique.

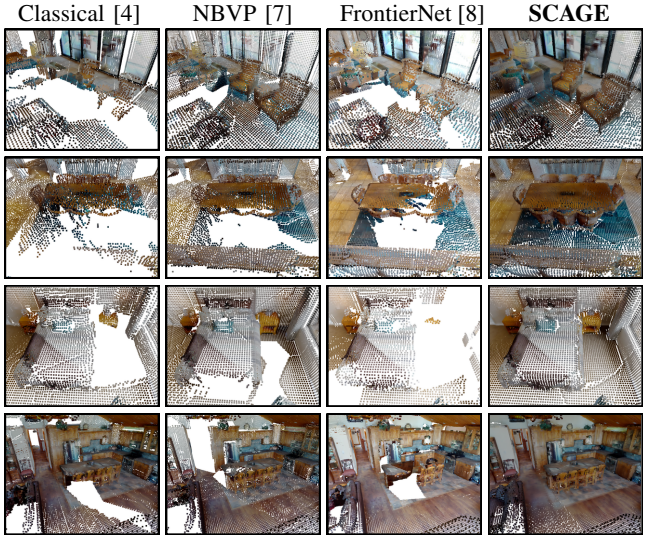

Fig. 5: **Qualitative Comparison. (Reconstruction quality).** Results comparing reconstruction quality of our method against baselines. By not just focusing on maximizing the coverage, but capturing all the structural details of indoor scenes, our method achieves the best visual fidelity. We are able to reconstruct *completely* all elements that compose a house: not just chairs, dining tables and beds but floor and ceiling as well; while other methods are only able to achieve partial reconstruction. Notice the white areas that represent regions of the scene which these methods fail to scan and reconstruct.

error using perfect depth and about 60% using depth estimated by Metric3D v2 [32]). Furthermore, SCAGE exhibits a remarkably low standard deviation across multiple evaluation runs with varying starting positions. This highlights the inherent robustness of the anomaly-driven formulation, proving that the system reliably captures high-fidelity maps regardless of the robot's initial initialization pose.

Qualitative Results: Coverage and Exploration Trajectories. To visually demonstrate the behavioral differences between the exploration strategies, we map the resulting trajectories and coverage across three distinct environments (Scenes 804, 812, and 824) in Fig. 4. Starting from identical initial locations, baseline methods frequently stall or leave substantial portions of the environment unexplored.

Notably, while methods like FrontierNet [8] appear to achieve numerically competitive coverage, their qualitative behavior reveals a critical flaw. FrontierNet tends to maximize its volumetric reward by simply visiting the entrances of rooms, failing to penetrate deeper into the space to observe the structural elements that actually compose the room (see Fig. 4). In contrast, by modeling exploration as a geometric anomaly minimization problem, SCAGE actively seeks out unresolved geometries, driving the robot deep into rooms to capture fine details.

Qualitative Results: Reconstruction Quality. This targeted resolution of structural anomalies translates directly into vastly superior visual fidelity, as highlighted in Fig. 5. While traditional methods typically yield fragmented or partial reconstructions of occluded objects, SCAGE captures the complete structural details of the indoor scenes. Our anomaly-guided approach successfully and completely reconstructs a diverse array of environmental elements. This includes complex, cluttered objects such as chairs, dining tables, and beds, as well as broad, featureless surfaces like floors and ceilings. By prioritizing structural completeness rather than sheer spatial expansion, our method ensures the final 3D map is both volumetrically expansive and geometrically accurate.

Ablation Study: Utility Function Design. We ablate the components of our utility function (Eq. 4) in Table II. Relying solely on the anomaly score ($\bar{s}_k$) yields the second-highest coverage, highlighting the raw exploratory power of targeting structural ambiguities. Penalizing it by distance ($\bar{s}_k / d(T_k)$) degrades performance, as the traversal term biases selection toward nearby targets and causes the robot to over-commit locally. Dropping the anomaly signal entirely ($g(T_k)/d(T_k)$, equivalent to NBVP [7]) is weakest of all: with no structural cue to distinguish which unobserved

Utility Function	Scene ID			Mean
	804	876	834	
$\bar{s}_k$	77.0 ± 6.5	74.0 ± 7.8	73.1 ± 9.2	74.7
$\bar{s}_k / d(T_k)$	73.5 ± 6.2	69.9 ± 8.6	72.0 ± 9.8	71.8
$g(T_k) / d(T_k)$	38.1 ± 6.2	63.4 ± 9.5	63.1 ± 9.1	54.9
$g(T_k) \cdot \bar{s}_k / d(T_k)$ (Ours)	92.5 ± 1.0	91.0 ± 0.1	84.1 ± 0.5	89.2

TABLE II: **Ablation Study: Utility Function Design.** Volumetric coverage (%) across three scenes comparing different utility formulations. $g(T_k)$ denotes volumetric information gain, $\bar{s}_k$ the mean cluster anomaly score, and $d(T_k) = \max(\|\mathcal{P}_{\text{robot}} - \mathcal{P}_{\text{cam},k}\|_2, \epsilon)$ the traversal cost. The results demonstrate that integrating all three terms is essential for maximizing exploration performance.

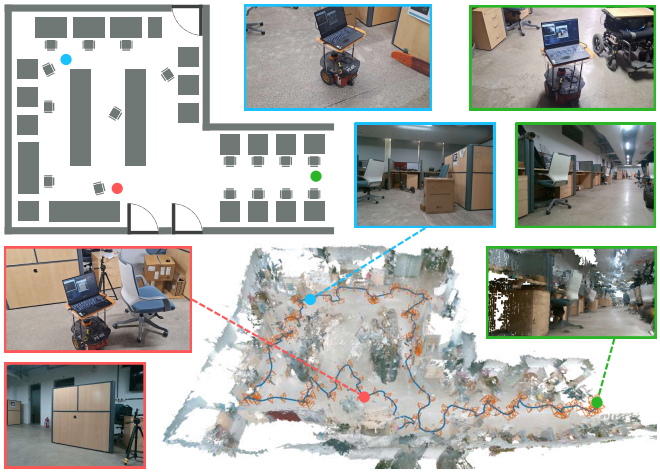

Fig. 6: **Real world validation**. Schematic representative map of the lab (top-left) and the resulting 3D point cloud reconstruction (roof cropped for visibility). The robot's trajectory begins at the green starting point. The surrounding color-coded images corresponding to the green, blue, and red markers, display the robot in the environment alongside its egocentric view at that exact moment during capture, including a point cloud from the starting location.

regions are worth resolving, the robot settles for whichever nearby free space is cheapest to reach, most visibly in complex scenes such as 804. The terms are thus complementary, $\bar{s}_k$ identifying where geometry is unresolved and $g(T_k)$ pulling the robot toward large unobserved volumes, counteracting the local bias of $d(T_k)$. Only their combination unlocks peak exploration behavior.

C. Real-world Validation

We implement SCAGE as a ROS package and deploy it on a P3DX robot equipped with a RealSense D455 depth camera, providing 640×360 RGB-D images at 10 Hz. All computations are performed on a laptop with an i5-9300H CPU, 32 GB of RAM, and an RTX 2060 GPU, achieving ∼7 Hz inference for real-time point cloud processing. We show the results in Fig. 6. Despite being trained on perfect-looking houses from the Matterport3D dataset, SCAGE is robust in the real world environment, with the robot successfully mapping the structural components of the large indoor scene. Notably, Δ-Net flags incomplete chairs, cubicle partitions, and walls in our lab as anomalies despite these specific instances differing from those seen during training, confirming that the network has learned the underlying geometry

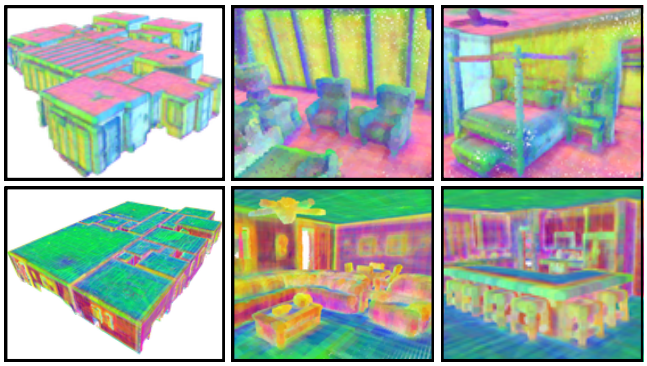

Fig. 7: **Emergent Structural Understanding.** PCA visualization of features extracted from the Δ-Net encoder on scenes 876 (top) and 879 (bottom). Structurally similar regions share consistent feature representations (colors) across different rooms. For example, planar horizontal surfaces like floors, ceilings, and beds exhibit uniform color, while objects such as fans, chairs, and kitchen cabinetry stand out in sharp contrast against their structural background.

of what constitutes such structures rather than memorizing training exemplars. However object categories absent from the training distribution, such as other mobile robots in the lab, elicit weak anomaly responses and are observed via anomalies of the surrounding area.

D. Emergent Structural Understanding

Beyond quantitative coverage and reconstruction metrics, we observe that the success of SCAGE is rooted in the network's deep, intrinsic understanding of indoor environments. To investigate what the network learns, we extract the dense latent features from the Δ-Net encoder and project them into RGB space using Principal Component Analysis (PCA) (Fig. 7). Even though the network is trained entirely from scratch without any explicit semantic supervision, it naturally learns to group similar structural and semantic elements. Horizontal surfaces (floors, ceilings, and beds) share consistent feature distributions, while distinct architectural objects like ceiling fans, chairs, and tables are highly isolated. This emergent structural awareness highlights the remarkable expressivity of the Point Transformer V3 [13] backbone. Because the model genuinely learns the fundamental building blocks of how houses are constructed, it can accurately identify *scene anomalies* to drive exploration. Furthermore, the richness of these learned representations suggests that our pre-trained encoder could serve as a powerful, off-the-shelf feature extractor for complex downstream applications, such as zero-shot semantic segmentation or object goal navigation.

V. CONCLUSION

We presented SCAGE, a novel autonomous exploration framework that successfully bridges the gap between spatial coverage and high-fidelity 3D reconstruction. By reframing exploration as a geometric anomaly-minimization problem, we move beyond traditional heuristics for exploration. At the core of our approach is Δ-Net, a 3D denoising autoencoder that learns the foundational structural priors of typical indoor environments directly from localized point cloud chunks.

During navigation, our system continuously evaluates live 3D observations against these learned expectations, identifying geometric contradictions as *scene anomalies*. By dynamically targeting and resolving these structural ambiguities from optimal vantage points, SCAGE seamlessly couples spatial discovery with precise geometric mapping. Extensive evaluations demonstrate that our approach achieves superior volumetric coverage ($\sim$90% across scenes, $\sim$15% improvement) and higher reconstruction quality (half the error) compared to state-of-the-art baselines. Ultimately, by shifting the primary exploratory signal from simply finding unmapped voids to actively achieving structural comprehension, this work establishes a new direction for autonomous 3D scene exploration.

Acknowledgements. The authors acknowledge the support provided by MeitY, Govt. of India, under the project "Capacity Building for Human Resource Development in Unmanned Aircraft System (Drone and Related Technology)."

REFERENCES

[1] M. Tranzatto, F. Mascarich, L. Bernreiter, C. Godinho, M. Camurri, S. Khattak, T. Dang, V. Reijgwart, J. Löje, D. Wisth *et al.*, "Cerberus: Autonomous legged and aerial robotic exploration in the tunnel and urban circuits of the darpa subterranean challenge," *Field Robotics*, vol. 2, pp. 274–324, 2022.

[2] C. Gao, F. Daxinger, L. Roth, F. Maffra, P. Beardsley, M. Chli, and L. Teixeira, "Aerial image-based inter-day registration for precision agriculture," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 862–11 868.

[3] M. F. Ginting, D. D. Fan, S.-K. Kim, M. J. Kochenderfer, and A.-A. Agha-Mohammadi, "Semantic belief behavior graph: Enabling autonomous robot inspection in unknown environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7604–7610.

[4] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

[5] W. Gao, M. Booker, A. Adiwahono, M. Yuan, J. Wang, and Y. W. Yun, "An improved frontier-based approach for autonomous exploration," in *2018 15th international conference on control, automation, robotics and vision (ICARCV)*. IEEE, 2018, pp. 292–297.

[6] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon" next-best-view" planner for 3d exploration," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1462–1468.

[7] 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.

[8] B. Sun, H. Chen, S. Leutenegger, C. Cadena, M. Pollefeys, and H. Blum, "Frontiernet: Learning visual cues to explore," *IEEE Robotics and Automation Letters*, vol. 10, no. 7, pp. 6576–6583, 2025.

[9] S. Upadhyay, K. M. Krishna, and S. Kumar, "Fast frontier detection in indoor environment for monocular slam," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, 2016, pp. 1–8.

[10] Y. Tao, E. Iceland, B. Li, E. Zwecher, U. Heinemann, A. Cohen, A. Avni, O. Gal, A. Barel, and V. Kumar, "Learning to explore indoor environments using autonomous micro aerial vehicles," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 758–15 764.

[11] L. Schmid, M. N. Cheema, V. Reijgwart, R. Siegwart, F. Tombari, and C. Cadena, "Sc-explorer: Incremental 3d scene completion for safe and efficient exploration mapping and planning," *arXiv preprint arXiv:2208.08307*, 2022.

[12] Y. Tao, Y. Wu, B. Li, F. Cladera, A. Zhou, D. Thakur, and V. Kumar, "Seer: Safe efficient exploration for aerial robots using learning to predict information gain," *arXiv preprint arXiv:2209.11034*, 2022.

[13] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler faster stronger," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 4840–4851.

[14] A. Dai, S. Papatheodorou, N. Funk, D. Tzoumanikas, and S. Leutenegger, "Fast frontier-based information-driven autonomous exploration with an mav," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 9570–9576.

[15] M. Selin, M. Tiger, D. Duberg, F. Heintz, and P. Jensfelt, "Efficient autonomous exploration planning of large-scale 3-d environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1699–1706, 2019.

[16] C. Cao, H. Zhu, H. Choset, and J. Zhang, "Tare: A hierarchical framework for efficiently exploring complex 3d environments." in *Robotics: Science and Systems*, vol. 5, 2021, p. 2.

[17] Z. Yan, H. Yang, and H. Zha, "Active neural mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 981–10 992.

[18] S. Lee, L. Chen, J. Wang, A. Liniger, S. Kumar, and F. Yu, "Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 070–12 077, 2022.

[19] W. Jiang, B. Lei, K. Ashton, and K. Daniilidis, "Ag-slam: Active gaussian splatting slam," *arXiv preprint arXiv:2410.17422*, 2024.

[20] W. Jiang, B. Lei, and K. Daniilidis, "Fisherrf: Active view selection and mapping with radiance fields using fisher information," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–440.

[21] Y. Tao, D. Ong, V. Murali, I. Spasojevic, P. Chaudhari, and V. Kumar, "Rt-guide: Real-time gaussian splatting for information-driven exploration," *IEEE Robotics and Automation Letters*, 2025.

[22] S. Papatheodorou, N. Funk, D. Tzoumanikas, C. Choi, B. Xu, and S. Leutenegger, "Finding things in the unknown: Semantic object-centric exploration with an mav," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3339–3345.

[23] O. Alama, A. Bhattacharya, H. He, S. Kim, Y. Qiu, W. Wang, C. Ho, N. Keetha, and S. Scherer, "Rayfronts: Open-set semantic ray frontiers for online scene understanding and exploration," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 5930–5937.

[24] S. Kim, O. Alama, D. Kurdydyk, J. Keller, N. Keetha, W. Wang, Y. Bisk, and S. Scherer, "Raven: Resilient aerial navigation via open-set semantic memory and behavior adaptation," *arXiv preprint arXiv:2509.23563*, 2025.

[25] D. S. Chaplot, E. Parisotto, and R. Salakhutdinov, "Active neural localization," *arXiv preprint arXiv:1801.08214*, 2018.

[26] D. S. Chaplot, M. Dalal, S. Gupta, J. Malik, and R. R. Salakhutdinov, "Seal: Self-supervised embodied active learning using exploration and 3d consistency," *Advances in neural information processing systems*, vol. 34, pp. 13 086–13 098, 2021.

[27] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 42–48.

[28] J. Chen, G. Li, S. Kumar, B. Ghanem, and F. Yu, "How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers," *arXiv preprint arXiv:2305.16925*, 2023.

[29] K. Qu, J. Tan, T. Zhang, F. Xia, C. Cadena, and M. Hutter, "Ippon: Common sense guided informative path planning for object goal navigation," *arXiv preprint arXiv:2410.19697*, 2024.

[30] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[31] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *The international journal of robotics research*, vol. 30, no. 7, pp. 846–894, 2011.

[32] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 579–10 596, 2024.