GenNBV: Generalizable Next-Best-View Policy for Active 3D Reconstruction

Xiao Chen[1,2]   Quanyi Li[1]   Tai Wang[1]   Tianfan Xue[1,2]   Jiangmiao Pang[1,*]

[1]Shanghai AI Laboratory   [2]The Chinese University of Hong Kong

{cx123,tfxue}@ie.cuhk.edu.hk   quanyili0057@gmail.com   {wangtai,pangjiangmiao}@pjlab.org.cn
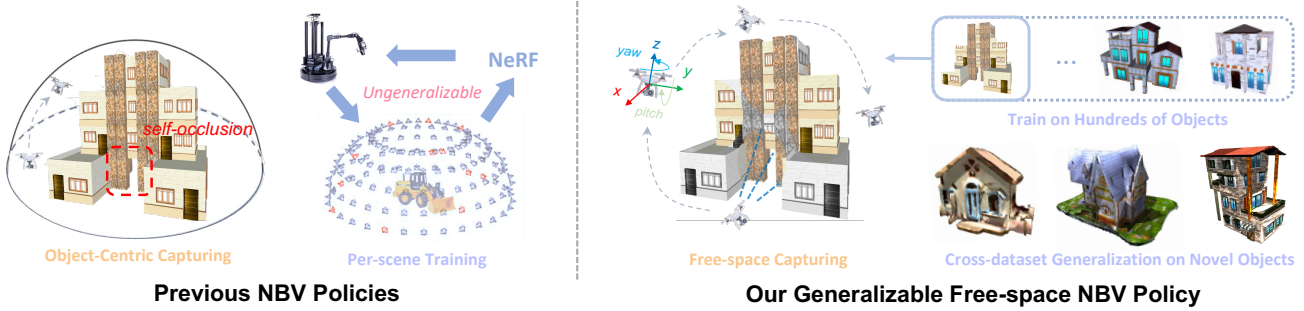
Figure 1. To determine the best view for 3D reconstruction, previous methods only chose from hand-crafted action space or based on object-centric capturing, lacking the ability to generalize to unforeseen scenes (Left). With our end-to-end trained, generalized free-space policy, it can generalize to unseen objects, enabling the captured drone to image from any viewpoint (Right).

Abstract

*While recent advances in neural radiance field enable realistic digitization for large-scale scenes, the image-capturing process is still time-consuming and labor-intensive. Previous works attempt to automate this process using the Next-Best-View (NBV) policy for active 3D reconstruction. However, the existing NBV policies heavily rely on hand-crafted criteria, limited action space, or per-scene optimized representations. These constraints limit their cross-dataset generalizability. To overcome them, we propose **GenNBV**, an end-to-end generalizable NBV policy. Our policy adopts a reinforcement learning (RL)-based framework and extends typical limited action space to 5D free space. It empowers our agent drone to scan from any viewpoint, and even interact with unseen geometries during training. To boost the cross-dataset generalizability, we also propose a novel multi-source state embedding, including geometric, semantic, and action representations. We establish a benchmark using the Isaac Gym simulator with the Houses3K and OmniObject3D datasets to evaluate this NBV policy. Experiments demonstrate that our policy achieves a 98.26% and 97.12% coverage ratio on unseen building-scale objects from these datasets, respectively, outperforming prior solutions.*

1. Introduction

Recent advances in 3D reconstruction [26, 40, 41] and neural rendering [23, 39, 50] have significantly enhanced the quality of 3D digitization of large scenes, such as buildings and city landmarks [12, 20, 21, 45, 49]. However, image-capturing process still remains time-consuming and labor-intensive. To scan a building-scale scene using a drone, it may take several days' effort of a professional team to ensure a full coverage. Moreover, even professional pilots may miss some areas at the first scan, leading to multiple rounds for rescanning.

To alleviate the effort in manual scanning, active 3D reconstruction algorithms, especially the Next-Best-View (NBV) policy, have emerged as a promising approach to automate view planning. However, typical rule-based NBV policies [10, 17, 25, 28, 30, 48] heavily rely on empirically designed view-selection criteria and hand-crafted action space, such as a hemisphere as shown in Fig. 1, which limit their generalizability to unseen scenes. To enhance generalization capacity, pioneering works on learning-based NBV policy have been proposed [3, 5, 28, 48]. Nevertheless, they are constrained into impractical setups in real-world scenarios, such as object-centric capturing and limited action space for ground robots, resulting in incomplete 3D reconstruction because of significant self-occlusion.

Therefore, in this work, we study this problem: *"How to design an NBV policy that supports exploration in any*

---

space and also can generalize to unseen building-scale geometries?" Building upon the design of previous reinforcement learning-based (RL-based) NBV policies [3, 5, 28], our setup faces the following new challenges that were not fully studied before: 1) Designing an easy-to-generalize action space, rather than object-specific spaces used in previous work, such as hand-crafted candidates or a constrained hemisphere; 2) With a large action space, an informative representation is necessary to guide the policy to efficiently find an optimal capturing trajectory; 3) Proposing generic reward functions and developing the distributed training procedure for better generalizability.

To overcome these two challenges, we propose *GenNBV*, an end-to-end generalizable NBV policy. It has a novel design of action spaces and state embeddings, which allows free space exploration and ensures a large coverage when applying to unseen objects. We firstly extend former limited action space, like a hemisphere, to 5D free space. As shown in Fig. 1, this new space is composed of a position subspace of approximately $20m$ x $20m$ x $10m$ and an omnidirectional heading subspace. This free space enables our agent drone to scan from any viewpoint. Moreover, with much larger action space, our RL-based policy can be easily generalized to building-scale geometries without concerning the potential scale shift from training to evaluation. On the contrary, many hand-crafted designs [17, 25] limit available viewpoints to a small predefined set, and thus cannot generalize if there is drastic change in scales.

Second, in order to generalize a well-trained NBV policy to unseen scenes during evaluation, we propose a novel state representation directly from raw sensory inputs to guide the policy. In contrast to the widely-used neural radiance field (NeRF) [17, 25, 30, 36, 48], our representation does not need time-consuming per-scene optimization. Specifically, our NBV policy incorporates a multi-source state embedding, which comprises three key components: a novel geometric embedding, a semantic embedding, and an action embedding. The geometric embedding is derived from multi-view depth maps and represents an encoded probabilistic 3D grid, while the semantic embedding is extracted from RGB images, and the action embedding is an encoded viewpoint sequence. In contrast to 2D representations used in previous RL-based NBV policies [3, 5], our multi-source state embedding, extracted from a probabilistic 3D grid, is a more comprehensive representation that supports a robust prediction of next viewpoints.

To validate the effectiveness of GenNBV, we construct a benchmark for training and evaluation with Houses3K [27] dataset from the NVIDIA Isaac Gym [22] simulator. We further test its generalization ability on novel buildings and object categories from OmniObject3D [44] and Objaverse [7] datasets. Our evaluation metrics quantify the completeness, accuracy, and efficiency for active 3D re-

construction. To consolidate completeness and efficiency into a single number, we also propose a metric, Area Under the Curve (AUC) of the coverage. Compared with heuristic [13], information gain-based [13, 17, 48], and RL-based baselines [28], our method achieves the best result under different metrics. We also provide visualizations to demonstrate the generalizability of GenNBV. A demo video is attached in the appendix to further illustrate our method.

2. Related Work

Traditional 3D Reconstruction. Photometry and geometry are two crucial aspects of reconstruction evaluation. Neural implicit representation [23, 26] has shown progress in photometric rendering performance but faces challenges like time-consuming optimization and poor generalization, hindering its application in real-time reconstruction problems such as 3D SLAM [14, 24, 32, 38, 40, 52]. Geometry, in contrast, typically corresponds to 3D representations such as point clouds and 3D mesh and is directly related to issues like collision avoidance. Considering these properties, researchers are prone to focus on geometric reconstruction in practical applications and we also follow this stream.

Active 3D Reconstruction. Active 3D reconstruction is a promising field that has not been thoroughly benchmarked yet. The pipeline of active 3D reconstruction frameworks alternates between inferring optimal viewpoints, capturing new data, and updating the rebuilt 3D model. With the optimal sequence of key viewpoints, a continuous planning path can be generated using the classic planners, such as RRT [16] and FMM [35].

Existing works can be pivotally differentiated based on the paradigm of NBV policies: rule-based or learning-based. Rule-based approaches like uncertainty-driven works [13, 17, 30, 48] typically yield the next best view from hand-crafted rules from the scene representation, which tends to overfit specific scenes. Most learning-based policies [3, 5, 28] use a deep reinforcement learning algorithm like PPO [34] to sequentially predict the optimal viewpoints based on observation. They must obtain feedback from task-related rewards such as coverage ratio [10, 28] and optimize during massive iteration with task environment.

The action space is designed based on the paradigm of NBV policies. Most rule-based NBV policies select views from a limited action space, such as a set composed of only one hundred candidate viewpoints [25], a tiny pre-defined space like a hemisphere [10, 17, 30, 48], making it possible to overlook some important viewpoints due to unavailability. Even though learning-based categories further explore the larger action space like a 2D plain [3, 5] or a constrained 3D space [28], their limited action space still prevents them from capturing sufficient details for 3D reconstruction.
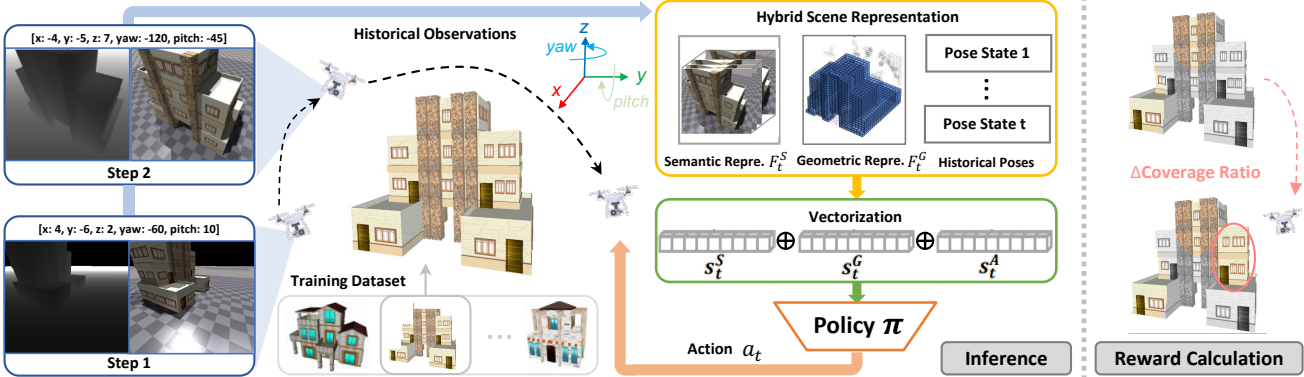
Figure 2. Overview of our proposed framework GenNBV. Our end-to-end policy takes the historical multi-source observations as input, transforms them into a more informative scene representation, and produces the next viewpoint position. A reward signal will be returned at training time to optimize the end-to-end policy for maximizing the expected cumulative reward in one episode. Specifically, the signal is the increased coverage ratio after collecting a new viewpoint.

Scene representation built from history observations, which directly provide reconstruction progress to NBV policies, is also a crucial aspect of the active 3D reconstruction framework. Previous works have explored visual representations for NBV policies, such as TSDF [12], neural radiance field [1], and 2D BEV maps [3, 11, 46]. However, implicit representations are hard to jointly with learning-based NBV policies, and 2D BEV map lacks sufficient information for large-scale outdoor scenes that contain numerous geometric details in 3D space.

In our benchmark, we select Scan-RL [28] as our main RL-based baseline. Compared to it, we introduce a much larger action space and informative visual representations for better generalizability. Additionally, we don't benchmark the mentioned baselines [3, 5], mainly due to their reliance on pretraining from human demonstration using imitation learning, while our focus lies on learning through iterations and end-to-end training. Different from our 3D free-space movement, they only explore the traversable 2D areas of indoor scenes using ground robots, thus limiting their motion space in 2D.

Generalizable Reinforcement Learning. Increasing the diversity of training data leads to better generalizability [6]. In the robotics field, domain randomization [43] allows legged robots to walk on various terrains absent from the training environment. For end-to-end driving policy, training in large-scale synthetic and realistic scenarios improves the safety of autonomous cars [18, 19] in unseen test scenarios. The advanced work [8] presents an RL-based formulation for the view planning problem. Based on similar frameworks of active 3D reconstruction, Scan-RL [28] and SCONE [10] are two pioneering works showing the generalizability of their frameworks, while both of them suffer from the constrain of viewpoint sampling in limited space and lack of in-depth analysis. In this work, we tackle

the aforementioned issues by predicting NBV in free 3D space with a learning-based planner, which is trained with a dataset containing buildings in various shapes and poses for acquiring generalizability for unseen buildings.

3. Methodology

In this section, we deliver our active 3D reconstruction framework GenNBV, especially the pivotal Next-Best-View policy. An overview of our framework is illustrated in Fig. 2. Firstly, we formulate the NBV problem as a Markov Decision Process (MDP), with a novel design of observations (blue boxes in Fig. 2) and action space in Sec. 3.1. Next, we elaborate our end-to-end NBV policy π (orange box in Fig. 2) in Sec. 3.2. Inspired by Curl [15], we point out that generalizability greatly depends on how to extract embeddings (green box in Fig. 2), which reflect the reconstruction progress, from raw sensory observations like RGB images and camera poses. In Sec. 3.3, we introduce the reward function (right-hand side of Fig. 2) reflecting the optimization objective and the details of policy optimization.

3.1. Formulation of the Next-Best-View Problem

We formulate the NBV problem as learning an optimal policy π that controls the capturing process, such that enough information is captured for large-scale scene reconstruction, with limited decision-making budgets. As capturing, transferring, and processing a large set of captured images introduce significant computational costs, we also want to design a policy that also minimizes the number of captured images. Therefore, our policy only captures sparse keyframes that sufficiently record all details of objects.

As shown in Fig. 2, at each time step t, the agent receives a visual observation o_t, takes an action a_t, infers the action at the next time step $t+1$ to move to a new location, and then receives a new visual observation, repeating this interaction

3

with the environment until the episode ends. Our simulated agent is embodied as CrazyFlie [9], a type of unmanned aerial vehicle equipped with various sensors, including an RGB-D camera and an IMU, to execute data collection for reconstruction. We discussed the details of the observation space and action space below.

Observation Space. As shown in the left column of Fig. 2, at each time step, the agent receives an RGB image I_t, a depth map D_t, and a state vector including the heading (yaw and pitch) and position (x, y, z) of the onboard camera. The observation o_t (yellow box in Fig. 2) consists of all previously captured images in one episode, historical actions, and the current captures and actions. With this information as input, the policy network can estimate the progress of data collection and determine where to scan next.

Action Space. Unlike most previous NBV algorithms, we use larger action space in order to cover all details of objects. Specifically, we use the camera location and camera angle as our action space, which is a 5-dimension vector consisting of 3D position coordinates and 2D rotation angles (yaw-axis and pitch-axis). We restrict the roll-axis rotation as it is not supported in all drone platforms.

3.2. Generalizable State Embedding

To ensure that the policy learned on one set of 3D objects can generalize to objects with different appearances and structures, a smart state embedding of raw sensor observations is needed to capture invariant features across different objects. Previous methods [3, 28] only extract representations in 2D space, which are very sensible to appearance changes, and our experiments have shown that policies only trained on 2D features may not generalize to different objects. Instead, we propose a multi-source representation that has better generalizability.

Specifically, we first build two mid-level representations that can better model the relationship between the 3D object and our agent: a 3D geometric representation F^G from depth maps and a semantic representation F^S from RGB images. Then we encode these mid-level representations and concatenate them with a pose embedding into state embedding s_t, guiding the NBV policy for subsequent decision-making. The overall encoder for state embedding is shown in Fig. 2, and details of each representation are discussed below.

Geometric Representation One simple way to record the geometry of a 3D object is using a binary 3D occupancy map [3], where the value of each cell in the 3D cube indicates whether a cell contains a 3D object or not. However, since the 3D representation is gradually updated with newly captured data, this simple binary occupancy map cannot differentiate a real unoccupied cell from an unscanned cell. An unscanned cell is a strong indicator that the agent should capture more data in this region, while no further scan is

needed for a real unoccupied cell.

Previous works [3, 11, 46] simplify the 3D scene representation into a 2D Bird's Eye View map for actively reconstructing indoor scenes. However, the 2D BEV map lacks sufficient information for our large-scale outdoor scenes that contain many geometric details in 3D space. In addition, their wheeled robotic platforms are constrained to freely scan and represent these outdoor scenes. Therefore, to model the scanning process in 3D free space, we employ the probabilistic 3D grid [42] as our geometric 3D representation, which records the probability of each 3D voxel being captured or not. Specifically, we first obtain a 3D point cloud in the world coordinate by back-projecting all 2D pixels to 3D points, using the depth map D_t, camera intrinsic parameters, and camera pose a_t. By voxelizing the obtained point cloud, we then build a 3D occupancy grid that explicitly indicates the binary state (occupied or free) in this 3D space. Subsequently, we represent this 3D grid as a probabilistic occupancy grid F_t^G and extend the state space of voxels with three states (occupied, free, unknown).

During each scanning (one episode in RL), at each step $t+1$, we update the probabilistic occupancy grid F_{t+1}^G based on the preceding grid F_t^G and current observation. Specifically, the grid is updated through Bresenham's line algorithm [2], which casts the ray path in 3D space between the camera viewpoint and the endpoints among the point cloud back-projected from depth D_{t+1}. Following the classical occupancy grid mapping algorithm [42], we have the log-odds formulation of occupancy probability:

$$\log \mathrm{Odd}(v_i|z_j) = \log \mathrm{Odd}(v_i) + C, \tag{1}$$

where v_i denotes the occupancy probability of ith voxel in the grid F_t^G, z_j is the measurement event that j^{th} camera ray passes through this voxel and C is an empirical constant. The derivation can be found in Appendix. Thus, we update the log-odds occupancy probability of each voxel in the grid F_t^G by adding a constant one time when a single camera ray passes through this voxel. Note that the probabilistic occupancy grid F^G is continually updated within one episode. Finally, the occupancy status of voxels is classified into three categories: unknown, occupied, and free, using preset probability thresholds.

Semantic Representation. Geometric representation enables agents to comprehend spatial occupancy, yet it's insufficient for perceiving the environment. For example, when observing a hole in an object, the agent may struggle to differentiate between incomplete scanning and the actual presence of a hole in the object. In such cases, the semantic information contained in the captured RGB images can help the agent distinguish between these two scenarios.

To provide semantic information, we employ a preprocessing module that takes as input the current frame of RGB image I_t and preceding k frames and converts these frames

$[I_t, I_{t-1}, ..., I_{t-k}]$ to grayscale, and concatenate them as output, following [28]. Then the preprocessed frames are fed into a two-layer convolutional network for extracting the semantic representation F_t^S.

State Embedding. To further combine the semantic and geometric embeddings, we first encode them to $s_t^S = f^S(F_t^S)$ and $s_t^G = f^G(F_t^G)$ where f^* are learnable networks $\mathrm{Linear}(\mathrm{Flatten}(x))$, as shown in Fig. 2. Subsequently, we combine them with the historical action embedding $s_t^A = \mathrm{Linear}(a_{1:t})$ to generate the final state embedding s_t, as the input to the policy network. This process can be formulated as: $s_t = \mathrm{Linear}(s_t^G; s_t^S; s_t^A)$.

Policy Network. Taking the state embedding s_t as input, the policy network is a 3-layer multi-layer perceptron network (MLP) whose output is used to parameterize a normal distribution over action space. In this way, the action can be drawn from the stochastic policy $a \sim \pi(\cdot|o_t)$.

3.3. Reward Function and Optimization

We train the end-to-end policy with reinforcement learning (RL) and hence design a reward function to precisely reflect the task objective for 3D reconstruction. The policy is optimized with proximal policy optimization [34] (PPO) for parallelizing sampling.

Reward Function. With the occupancy probability F_t^G at time step t, we can threshold each voxel with an empirical bound to determine if it is occupied. This discrimination process outputs a binary occupancy grid with $\tilde{N}_t$ voxels being occupied, which is used to calculate the coverage ratio:

$$\mathrm{CR}_t = \frac{\tilde{N}_t}{N^*} \cdot 100\%, \qquad (2)$$

where N^* is the number of ground-truth occupied voxels representing the surface of objects. To encourage our NBV policy to cover as many unseen areas of objects as possible, we use the difference of coverage ratio (CR) between two consecutive steps as the main reward function r^{CR}:

$$r_{t+1}^{\mathrm{CR}} = \mathrm{CR}_{t+1} - \mathrm{CR}_t. \qquad (3)$$

In free-space exploration, we also need to avoid collision. Previous limited-space agents [10, 17, 48] do not consider collision avoidance since their search space, like hemisphere, is by-design safe. Thus, we add a negative reward for collision and terminate the episode if a collision happens or the current coverage ratio is greater than 95%. We also implement a negative reward when the number of captured keyframes is over an empirical threshold to improve the path efficiency.

Policy Optimization. Once the reward function has been specified, the policy can be learned through any off-the-shelf RL algorithm. In this work, we specifically use PPO as thousands of workers can be parallelized to improve the

sample efficiency. Specifically, given our parameterized policy π_θ, the objective of PPO is to maximize the following objective function:

$$L(\theta) = \mathbb{E}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} A^{\pi_{\theta_{\mathrm{old}}}}(s_t, a_t) \right], \qquad (4)$$

where $A^{\pi_{\theta_{\mathrm{old}}}}(s_t, a_t)$ is the advantage function that measures the value of taking action a_t at state s_t under the current policy $\pi_{\theta_{\mathrm{old}}}$. To prevent significant deviation of the new policy from the old policy, PPO incorporates a clipped surrogate objective function:

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t[\min(\eta_t(\theta)A^{\pi_{\theta_{\mathrm{old}}}}(s_t, a_t),$$
$$\mathrm{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A^{\pi_{\theta_{\mathrm{old}}}}(s_t, a_t))], \qquad (5)$$

where $\eta_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$ and ϵ is a hyper-parameter that controls the size of the trust region.

4. Experiments

In this section, we conduct experiments on Houses3K [27], OmniObject3D [44], and Objaverse [7]. For our policy and other learning-based policies, we train them on Houses3K training datasets. After this, we evaluate all policies on the Houses3K test dataset and the OmniObject3D dataset for quantifying the in-distribution and out-of-distribution generalizability. Finally, we show the visualization results on three datasets, demonstrating the effectiveness of the proposed method.

4.1. Experimental Setup

Simulation Environment. We conduct all experiments in NVIDIA Isaac Gym [22], a physics simulation platform designed for reinforcement learning and robotics research. With GPU-accelerated tensor API and sensor interaction API, we can easily implement customized CUDA kernels to calculate our geometric representation efficiently. Moreover, the sensor simulation is efficient. It can run up to 1000 FPS, significantly reducing the training time. We create the agent drone CrazyFlie [9] and equip it with sensors including RGB-D cameras and IMU. Considering memory limitations, we downsample the image resolution to 400×400. The vertical field of view of this onboard camera is 90°. We also set up a point light source with a fixed position and constant intensity in Isaac Gym.

Dataset. We conduct our experiments on Houses3K [27], OmniObject3D [44] and Objaverse 1.0 [7]. Our model is trained on large-scale 3D objects from Houses3K. To validate the effectiveness and generalizability of our model, we evaluate it on *unseen* 3D objects from Houses3K. For testing the cross-dataset generalizability, we also evaluate it on batches of various objects from OmniObject3D and Objaverse datasets, which include both house and non-house categories.

5

Table 1. Evaluation results of Next-Best-View policies for active 3D reconstruction on **Houses3K** and the house category from **OmniObject3D** (cross-dataset generalization). The number of views is set to 30 and 20 for Houses3K and OmniObject3D, respectively.
"*": the policy is trained with the Houses3K training set and evaluated with holdout Houses3K test set and OmniObject3D house category.
"†": the policy heavily relies on optimized per-scene representation (NeRF), and thus is directly trained and evaluated on testing objects.

	NBV Policy	Houses3K Test Set			OmniObject3D		
		Mean AUC ↑	Final Coverage Ratio ↑	Accuracy (cm) ↓	Mean AUC ↑	Final Coverage Ratio ↑	Accuracy (cm) ↓
Heuristic	Random	48.53%	58.24%	1.38	60.13%	73.73%	0.62
	Random Hemisphere	71.19%	79.72%	0.63	74.03%	84.74%	0.48
	Uniform Hemisphere	82.91%	89.71%	0.44	83.09%	92.90%	0.41
InfoGain	†Uncertainty-Guided [17]	83.13%	89.31%	0.44	69.08%	92.85%	0.42
	†ActiveRMap [48]	84.86%	90.77%	0.44	69.47%	93.15%	0.42
RL-based	*Scan-RL [28]	84.49%	91.61%	0.40	82.17%	92.53%	0.34
	*Scan-RL w/ Our Repre.	86.48%	92.20%	**0.37**	86.14%	93.73%	0.35
	*Ours w/ Scan-RL Repre.	87.39%	95.63%	0.38	86.81%	93.64%	0.34
	***GenNBV (Ours)**	**91.19%**	**98.26%**	**0.37**	**88.63%**	**97.12%**	**0.33**

Table 2. The cross-dataset generalization for **non-house categories**. We train the baseline Scan-RL and our GenNBV on Houses3K and generalize them to non-house categories from OmniObject3D and an indoor scene from Replica.

Category	Dataset	NBV Policy	AUC ↑	FCR ↑	Acc. ↓
Animals	OmniObject3D	Scan-RL	76.43%	84.98%	0.17
		GenNBV	**84.28%**	**94.07%**	**0.16**
Trucks	OmniObject3D	Scan-RL	69.58%	76.37%	0.06
		GenNBV	**75.54%**	**83.47%**	**0.03**
Dinosaurs	OmniObject3D	Scan-RL	78.47%	86.89%	0.88
		GenNBV	**84.24%**	**94.03%**	**0.81**
Room 0	Replica	Scan-RL	69.21%	80.75%	**3.12**
		GenNBV	**88.53%**	**99.16%**	3.59

Houses3K contains 3,000 textured 3D building models. They are divided into 12 batches, with each batch featuring 50 distinct geometries and 5 varying textures. We further eliminate poorly designed geometries based on the following two criteria: 1) objects with complex internal structures (as we are mostly focusing on the surface reconstruction), and 2) objects with redundant bases at the bottom. Following these criteria, the training set consists of 256 objects from 6 selected batches, and the test set consists of 50 objects from another batch. All these selected training and test objects have distinct geometries.

OmniObject3D offers a collection of 6K high-fidelity objects scanned from real-world sources across 190 typical categories. The house category from OmniObject3D for evaluation has 43 diverse objects. In contrast, Objaverse 1.0

comprises a vast library of 818K 3D synthetic objects spanning 21K categories. The diverse and high-quality nature of these datasets makes them ideal for evaluating our models and visualizing the results.

Evaluation Metrics. The objective of NBV policies is to capture the most useful information for reconstruction, with the least number of views. We report the *(1) Final Coverage Ratio (%)*, *(2) Mean AUC (%)* of coverage ratio and *(3) Reconstruction Accuracy (cm)* along with the consistent number of views in all tables. Specifically, most prior works [10, 28] separately use the coverage ratio (%) and the number of views to evaluate the reconstruction completeness and efficiency for NBV policies. However, the coverage ratio is highly correlated with the number of views. Thus we propose to unify the number of views to a fixed value for all NBV policies during evaluation and use the area under the curve (AUC) of coverage ratio as the main metric for comparison. Also, we calculate the Chamfer Distance between scanned and ground-truth point clouds to represent the reconstruction accuracy (cm).

Implementation Details. We conduct all experiments in Isaac Gym simulation engine with one NVIDIA Tesla V100 GPU. Our implementation refers to the codebase of Legged Gym [33] and the PPO implementation in Stable Baseline3 [29]. The ground-truth point clouds on objects' surfaces are sampled by the Poisson Disk sampling method [47] using Open3D API [51]. The Chamfer Distance between the scanned point cloud and the ground-truth point cloud is calculated using PyTorch3D API [31]. Please refer to the Appendix for further details.

(a) Houses3K **(b) OmniObject3D**

Figure 3. The visualization results of unseen 3D objects reconstructed by Scan-RL [28] and our model to compare the generalizability. (a) Unseen buildings from the test set of Houses3K. (b) Unseen buildings from OmniObject3D. It's quite obvious that some parts of the model reconstructed by Scan-RL are wrong or missing. For example, the second object in the first row has a pillar in the wrong shape. Scan-RL fails to reconstruct the roof edge for the fourth object from OmniObject3D, as shown in the third row.

4.2. Performance Comparison

To comprehensively demonstrate the effectiveness and generalizability of GenNBV, we design three levels of evaluation experiments: 1) As shown in Table 1, we show the performance of our NBV policy on the Houses3K test set; 2) We generalize the policy trained on Houses3K to the house category from OmniObject3D that has completely different geometric structures and textures compared to the training set. The quantitative result is at Table 1 and the visualization result is as shown in Fig. 3; 3) We also generalize GenNBV to non-house categories from OmniObject3D and scenes with enormous details from Objaverse and Replica [37] to demonstrate its generalizability potential.

We implement the following policies as our baselines. 1) **Random**: This policy randomly generates 5-dim vector $(x, y, z, pitch, yaw)$ among the action space as the next viewpoint. 2) **Random Hemisphere**: This policy randomly generates the next positions on a pre-defined hemisphere that sufficiently covers all objects of the test set. The headings are constrained to point to the center of the hemisphere. 3) **Uniform Hemisphere**: All positions are evenly distributed on the previously mentioned hemispheres. 4) **Uncertainty-Guided Policy** [17]: This NBV policy iteratively selects the next view from a pre-defined viewpoint set according to the uncertainty based on a continually optimized neural radiance field. 5) **ActiveRMap** [48]: This policy also adopts an iterative selection framework, with multiple objectives including information gain, to select the next viewpoints from the candidate set. 6) **Scan-RL** [28]: This RL-based policy predicts the next viewpoint from a hemisphere space, relying on the historical RGB images. 7) **Ours with Scan-RL's Representation**: This free-space NBV policy replaces our representation with Scan-RL's representation only extracted from RGB images.

Table 3. Ablation studies of representation categories in our framework on unseen Houses3K test set.

	Representation Category			Evaluation Metrics	
	Probabilistic 3D Grid	5-DoF Pose	Semantic 2D Map	Mean AUC	Final Coverage Ratio
Uni-source	✓			81.06%	84.56%
		✓		69.53%	76.61%
			✓	81.24%	87.90%
Multi-source	✓	✓		88.66%	96.67%
	✓		✓	89.77%	95.31%
		✓	✓	88.30%	96.29%
	✓	✓	✓	**91.19%**	**98.26%**

As shown in Table 1, our GenNBV shows the best in-distribution and out-of-distribution generalizability in both coverage sufficiency and view efficiency, when evaluated on test sets consisting of unseen objects from Houses3K and OmniObject3D. In particular, the comparison in Table 1 demonstrates the strength of the proposed representations and action space. Note that the Scan-RL's representation ('*' in the table) uses 6 frames, while our semantic representation only needs 2 frames. Moreover, learning-based NBV policies such as GenNBV and Scan-RL, with much larger action space, outperform all rule-based baselines, even if ActiveRMap and Uncertainty-Guided baselines are trained and evaluated on the same 3D objects.

Non-house Generalization In Table 2, we introduce an additional experiment to evaluate the generalizability of our GenNBV trained on Houses3K to non-house categories. The targets include 74 *animals*, 43 *trucks* and 33 *dinosaurs* from OmniObject3D, and a challenging *indoor*

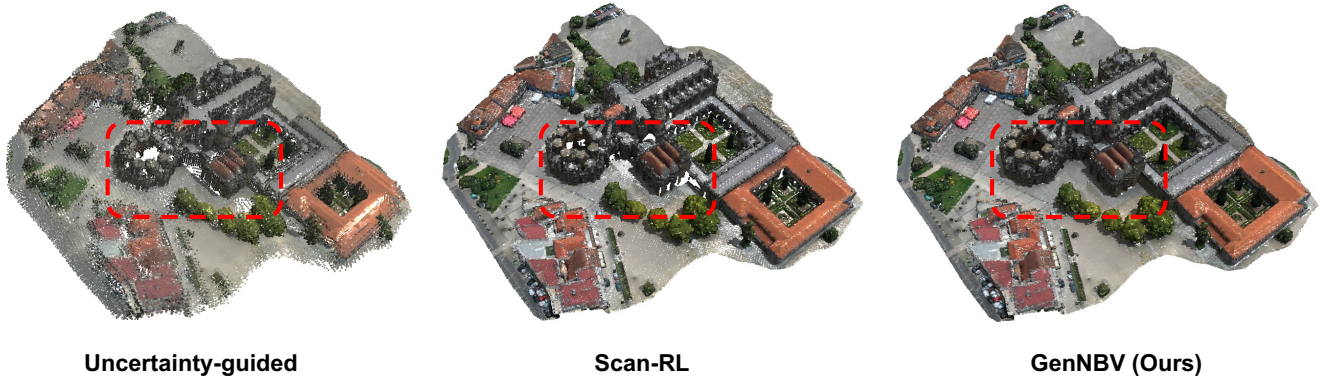**Uncertainty-guided**　　　　　**Scan-RL**　　　　　**GenNBV (Ours)**

Figure 4. The visualization results of an unseen 3D outdoor scene with enormous details from Objaverse, reconstructed by Uncertainty-guided, Scan-RL and our model. Compared to the uncertainty-guided method and Scan-RL, the scene reconstructed by our method is more watertight and has fewer holes on the ground and building surface, especially in the region highlighted by the red box.

scene (Room 0) from Replica. The number of views is set to 20 for all experiments.

Our GenNBV almost outperforms Scan-RL in all metrics. Particularly, on Replica Room 0, which consists of numerous multi-sized objects with complex occlusion, our GenNBV is significantly better than Scan-RL. This shows the effectiveness of our free-space NBV policy.

4.3. Ablation Study

Multi-source State Embedding. In Sec. 3.2, we introduce how to build our multi-source state embedding. Thanks to the same shape of representations and state embedding, we only need to adjust the input dimension of the linear layer of state embedding according to the modal type when ablating the model. Here we reveal the importance of specific representations with the ablation results shown in Table 3. In uni-source experiments, we design the 5-DoF Pose baseline that learns an empirically optimal trajectory, which helps understand the effectiveness of domain knowledge of the dataset. We also evaluate different combinations of representation categories in Table 3, demonstrating the effectiveness of our multi-source state embedding for policy learning in multi-source experiments.

Table 4. Ablation studies of depth-based representations on unseen Houses3K test set, where depth map is the only sensory source.

Depth-based Representation	Mean AUC	Final Coverage Ratio
Geometric 2D Map	71.86%	74.88%
Binary 3D Grid	77.28%	80.85%
Probabilistic 3D Grid	**81.06%**	**84.56%**

Depth-based Representation. Furthermore, we ablate the depth-based representations in Table 4. The experimental results demonstrate that our probabilistic 3D grid is more

comprehensive than a binary 3D grid and a geometric 2D map to support the NBV prediction.

Number of Training Objects. Motivated by generalizable RL-based policies [6, 18, 19], we explore the impact of diversity of training data for generalizability. As shown in Table 5 and Fig. 5, we found that increasing the diversity of training objects indeed leads to better generalizability for 3D reconstruction.

Table 5. Ablation studies of the number of training objects in our framework on unseen OmniObject3D house category.

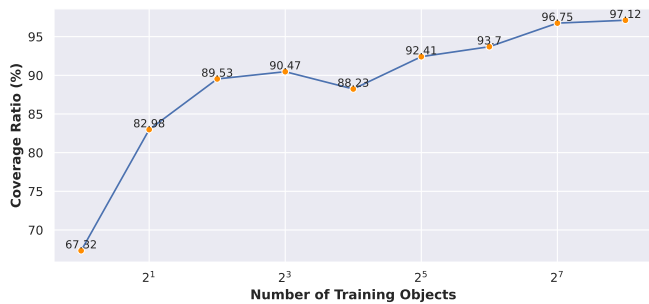| Number of Training Objects	Mean AUC	Final Coverage Ratio
1	61.76%	67.32%
2	71.03%	82.98%
4	70.72%	89.53%
8	73.71%	90.47%
16	73.73%	88.23%
32	82.39%	92.41%
64	82.09%	93.70%
128	87.21%	96.75%
256	**88.63%**	**97.12%**

Figure 5. The curve of coverage ratio with the increasing number of training objects on unseen OmniObject3D house category.

4.4. Qualitative Results

We visualize the reconstruction results generated from the scanning trajectory of a single episode in Fig. 3. It demonstrates that our next-best-view policy can reconstruct objects better in terms of completeness and appearance quality compared to Scan-RL. We further visualize the reconstruction results of an outdoor scene from Objaverse using 30 collected views in Fig. 4.

5. Conclusion

This study presents an end-to-end approach for active 3D scene reconstruction, reducing the need for manual intervention. Specifically, the learning-based policy explores how to reconstruct diverse objects in the training stage and thus can generalize to reconstruct unseen objects in a fully autonomous manner. Our controller maneuvers in free space and the next best view selection based on a hybrid scene representation which conveys scene coverage status and thus reconstruction progress. We show the effectiveness of our approach by testing it on multiple datasets including Houses3K, OmniObject3D, and Objaverse. The quantitative and qualitative generalization results on holdout Houses3K test set and cross-domain OmniObject3D house category show that our method outperforms other baselines in terms of reconstruction completeness, efficiency and accuracy. In addition, the experiment conducted on Objaverse shows that the policy trained in single-building settings can even generalize to a complicated outdoor scene.

References

[1] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2): 4606–4613, 2022. 3

[2] Jack E Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems journal*, 4(1):25–30, 1965. 4

[3] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 13

[5] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[6] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020. 3, 8

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 5, 13

[8] Mustafa Devrim Kaba, Mustafa Gokhan Uzunbas, and Ser Nam Lim. A reinforcement learning approach to the view planning problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6933–6941, 2017. 3

[9] Wojciech Giernacki, Mateusz Skwierczyński, Wojciech Witwicki, Paweł Wroński, and Piotr Kozierski. Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering. In *2017 22nd International Conference on Methods and Models in Automation and Robotics (MMAR)*, pages 37–42. IEEE, 2017. 4, 5

[10] Antoine Guedon, Pascal Monasse, and Vincent Lepetit. Scone: Surface coverage optimization in unknown environments by volumetric integration. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 5, 6

[11] Junfu Guo, Changhao Li, Xi Xia, Ruizhen Hu, and Ligang Liu. Asynchronous collaborative autoscanning with mode switching for multi-robot scene reconstruction. *ACM Transactions on Graphics (TOG)*, 41(6):1–13, 2022. 3, 4

[12] Guillaume Hardouin, Julien Moras, Fabio Morbidi, Julien Marzat, and El Mustapha Mouaddib. Next-best-view planning for surface reconstruction of large-scale 3d environments with multiple uavs. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1567–1574. IEEE, 2020. 1, 3

[13] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Con-*

ference on Robotics and Automation (ICRA), pages 3477–3484. IEEE, 2016. 2

[14] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17408–17419, 2023. 2

[15] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pages 5639–5650. PMLR, 2020. 3

[16] Steven M LaValle and James J Kuffner Jr. Randomized kinodynamic planning. *The international journal of robotics research*, 20(5):378–400, 2001. 2

[17] Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. *IEEE Robotics and Automation Letters*, 7(4):12070–12077, 2022. 1, 2, 5, 6, 7, 13

[18] Quanyi Li, Zhenghao Peng, Lan Feng, Chenda Duan, Wenjie Mo, Bolei Zhou, et al. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. In *Advances in Neural Information Processing Systems*, 2022. 3, 8

[19] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022. 3, 8

[20] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. *arXiv e-prints*, pages arXiv–2308, 2023. 1

[21] Yilin Liu, Liqiang Lin, Yue Hu, Ke Xie, Chi-Wing Fu, Hao Zhang, and Hui Huang. Learning reconstructability for drone aerial path planning. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022. 1

[22] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. 2, 5, 12

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 405–421. Springer, 2020. 1, 2

[24] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. *Robotics: Science and Systems*, 2022. 2

[25] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 230–246. Springer, 2022. 1, 2

[26] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1, 2

[27] Daryl Peralta, Joel Casimiro, Aldrin Michael Nilles, Justine Aletta Aguilar, Rowel Atienza, and Rhandley Cajote. Houses3k dataset. https://github.com/darylperalta/Houses3K, 2020. 2, 5, 13

[28] Daryl Peralta, Joel Casimiro, Aldrin Michael Nilles, Justine Aletta Aguilar, Rowel Atienza, and Rhandley Cajote. Next-best view policy for 3d reconstruction. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 558–573. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 13

[29] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *The Journal of Machine Learning Research*, 22(1):12348–12355, 2021. 6, 12

[30] Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. *IEEE Robotics and Automation Letters*, 2023. 1, 2

[31] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6

[32] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022. 2

[33] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *5th Annual Conference on Robot Learning*, 2021. 6

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 5

[35] James A Sethian. A fast marching level set method for monotonically advancing fronts. *proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996. 2

[36] Edward J Smith, Michal Drozdzal, Derek Nowrouzezahrai, David Meger, and Adriana Romero-Soriano. Uncertainty-driven active vision for implicit scene reconstruction. *arXiv preprint arXiv:2210.00978*, 2022. 2

[37] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 7

[38] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 2

[39] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields

reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 1

[40] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1, 2

[41] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3D reconstruction in the wild. In *SIGGRAPH Conference Proceedings*, 2022. 1

[42] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002. 4, 12

[43] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 3

[44] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2, 5, 13

[45] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *The European Conference on Computer Vision (ECCV)*, 2022. 1

[46] Kai Ye, Siyan Dong, Qingnan Fan, He Wang, Li Yi, Fei Xia, Jue Wang, and Baoquan Chen. Multi-robot active mapping via neural bipartite graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14839–14848, 2022. 3, 4

[47] Cem Yuksel. Sample elimination for generating poisson disk sample sets. *Computer Graphics Forum*, 34(2):25–32, 2015. 6, 13

[48] Huangying Zhan, Jiyang Zheng, Yi Xu, Ian Reid, and Hamid Rezatofighi. Activermap: Radiance field for active mapping and planning. *arXiv preprint arXiv:2211.12656*, 2022. 1, 2, 5, 6, 7, 13

[49] Han Zhang, Yucong Yao, Ke Xie, Chi-Wing Fu, Hao Zhang, and Hui Huang. Continuous aerial path planning for 3d urban scene reconstruction. *ACM Trans. Graph.*, 40(6):225–1, 2021. 1

[50] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022. 1

[51] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 6, 13

[52] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 2

GenNBV: Generalizable Next-Best-View Policy for Active 3D Reconstruction

Supplementary Material

A. Implementation Details

A.1. Occupancy Grid Mapping Algorithm

Before updating the probabilistic occupancy grid F_t^G, Bresenham's line algorithm is implemented to cast the ray path in 3D space between the camera viewpoint and the endpoints among the point cloud back-projected from D_{t+1}. According to the classical occupancy grid mapping algorithm [42], we have the log-odds formulation of occupancy probability:

$$\log Odd(v_i|z_j) = \log Odd(v_i) + \log \frac{p(z_j|v_i = 1)}{p(z_j|v_i = 0)}, \quad (6)$$

where v_i denotes the occupancy probability of ith voxel in the grid F_t^G, z_j is the measurement event that j^{th} camera ray passes through this voxel. The numerator of the item $\log \frac{p(z_j|v_i=1)}{p(z_j|v_i=0)}$ means that the probability of being passed for a voxel if this voxel is occupied in fact, which shows the confidence of ray casting process. Obviously, the numerator and denominator can be set as an empirical constant. Therefore, we update the log-odds occupancy probability of each voxel in the grid F_t^G by adding a constant one time when a single camera ray passes through this voxel. Note that the probabilistic occupancy grid F^G is continually updated within one episode. Finally, the occupancy status of voxels can be classified into three categories: unknown, occupied, and free, by setting a probability threshold.

We use a constant C' to represent the value of a ratio $|\frac{\log p(z_j|v_i=1)}{\log p(z_j|v_i=0)}|$ in Eq. (6), where the denominator term $|\log p(z_j|v_i = 0)|$ is set to 0.01. To demonstrate the robustness and optimality of hyper-parameter in our implementation, we show the experimental results in Tab. 6.

We implement this algorithm with PyTorch. In particular, we used ray-casting renderers from PyTorch3D [22] to generate and use depth maps as inputs to our model.

A.2. Network and Training

Training. In Isaac Gym [22], we set 256 parallel environments for policy training, where each environment corresponds with one building-level object. Once the coverage ratio reaches a threshold (99% in practice), a collision happens or the episode length reaches the maximum threshold (100 steps), the environment is reset and the building to be reconstructed is replaced. Our NBV policy is optimized through over 32 million iterations and using approximately 24-hour training time on a NVIDIA V100 GPU. All networks are randomly initialized and trained end-to-end.

Table 6. Evaluation results for different empirical parameters in the implementation of occupancy grid mapping algorithm. Note that we use a constant C' to represent the value of a ratio $|\log \frac{p(z_j|v_i=1)}{p(z_j|v_i=0)}|$ in Eq. (6)

Occupancy Threshold	Mean AUC ↑	Final Coverage Ratio ↑	Accuracy ↓
0.5 (Ours)	**87.39%**	**95.63%**	**0.37**
1.0	84.41%	93.43%	0.41
1.5	84.84%	95.20%	0.39
2.5	54.77%	58.70%	0.85
Value of Constant C'	**Mean AUC ↑**	**Final Coverage Ratio ↑**	**Accuracy ↓**
5	66.96%	71.92%	0.49
10	80.75%	92.16%	0.43
20 (Ours)	**87.39%**	**95.63%**	**0.37**
40	86.97%	94.12%	0.40

Network. We propose a multi-source representation that has better generalizability. In particular, we first build two mid-level representations: a 3D geometric representation F^G from depth maps and a semantic representation F^S from RGB images. Geometric representation F^G encoded from depth images is with shape of $(N, X, Y, Z, 4)$, where the number of parallel training environments N is 256, grid size (X, Y, Z) is $(20, 20, 20)$ and last dimension represents the 3D world coordinate and occupancy possibility. Semantic representation F_S is stacked grayscale images encoded from multi-view RGB images with the shape of (N, M, H, W), where the number of RGB images M is 5, the size of grayscale images is $(64, 64)$

Specifically, we encode mid-level representations to embeddings: $s_t^S = f^S(F_t^S)$ and $s_t^G = f^G(F_t^G)$. f^S encompasses a 2-layer 2D convolution and $\text{Linear}(\text{Flatten}(x))$ operation, while f^G encompasses a 2-layer 3D convolution and $\text{Linear}(\text{Flatten}(x))$. Subsequently, we combine them with the historical action embedding $s_t^A = \text{Linear}(a_{1:t})$ to generate the final state embedding s_t, as the input to the policy network. This process can be formulated as:

$$s_t = \text{Linear}(s_t^G; s_t^S; s_t^A), \quad (7)$$

where all embeddings are 256 vectors.

Our policy network PPO, implemented by Stable Baseline3 [29], is a 3-layer multi-layer perceptron network (MLP). The output of our policy is used to parameterize a distribution over our 5-dimension action space. In this way, the action can be drawn from the stochastic policy $a \sim \pi(\cdot|o_t)$.

A.3. Baseline Policies

The implementation details of baseline policies are described below:

1) **Random Policy**: This policy randomly generates 5-dim vector $(x, y, z, pitch, yaw)$ among the action space as the next action. The randomly generated positions are constrained so as not to cause collisions. The reported results are evaluated on the test set and averaged over random seeds from 0 to 4.

2) **Random Policy on the Sphere**: This policy randomly generates positions (x, y, z) on a pre-defined hemisphere that exactly covers all objects of the test set. The headings are required to point to the center of the hemisphere. To avoid collisions, we set the radius of the hemisphere to 9 meters, which is greater than the maximum height of the test set object. The reported results are evaluated on the test set and averaged over random seeds from 0 to 4.

3) **Uniform Policy on the Sphere**: All positions are evenly distributed on the previously mentioned hemispheres. Specifically, for Houses3K test set, all sampling points are distributed over 5 heights, each with 6 evenly spaced positions. For OmniObject3D, all sampling points are distributed over 4 heights, each with 5 evenly spaced positions.

4) **Uncertainty-Guided** [17]: This NBV policy iteratively selects the next view from a pre-defined viewpoint set based on the entropy-based uncertainty based on a continually optimized neural radiance field. We use TensoRF [4] as the implementation foundation of neural radiance field. Before implementing uncertainty-driven viewpoint selection, we uniformly sample 100 views on the pre-defined hemisphere as the viewpoint candidate set. Note that we need to sample the candidate set for each object.

5) **ActiveRMap** [48]: The working pipeline is similar to uncertainty-guided policy. In particular, we implement the "discrete (free)" setup of ActiveRMap, which constrains the drone agent on the pre-defined hemisphere. Having considered collision avoidance in the design of viewpoint set, we remove the collision penalty in its optimization objective.

6) **Scan-RL** [28]: This RL-based policy predicts the next viewpoint from a 3-dim hemisphere space, relying on the historical RGB.

7) **Ours with Scan-RL's Representation**: To further compare the former Next-Best-View policy Scan-RL [28] with us, we implement Scan-RL in our experimental setup, with our action and state space. This free-space NBV policy uses Scan-RL's representation only extracted from RGB images instead of our original hybrid representations.

B. Data Preprocessing

3D Mesh. To boost our NBV policy's generalizability on building-scale objects, we rescale the original 3D meshes from Houses3K [27], OmniObject3D [44] and Objaverse [7] to a reasonable building size. For example, the size of building meshes from Houses3K and OmniObject3D are approximately $(15m, 15m, 8m)$.

Ground-truth Point Cloud. We use the Poisson Disk sampling method [47] implemented by Open3D [51] to sample $100,000$ points from 3D meshes. And then we voxelize these point clouds. These voxelized points are viewed as ground-truth point clouds of these meshes.

C. Additional Experiments

The evidence for sufficient and efficient capturing. To evaluate both completeness and efficiency, we use the area under the curve (AUC) of coverage ratio as our main metric, stated in Sec. 4.1. Additionally, the figure below shows the mean AUC curves for OmniObject3D houses. Our GenNBV achieves a better coverage ratio under 20 views (97.12% v.s. 92.53%), and even surpasses the saturated coverage of Scan-RL using merely 5 views.
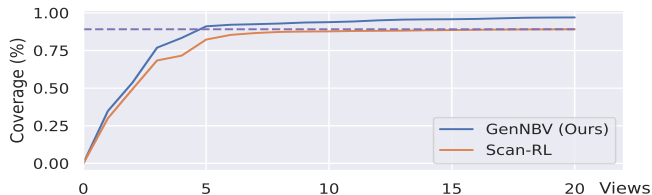
The effectiveness of 2D semantic representations
Below, we add the experiments on Houses3K test set to illustrate the effectiveness of the proposed semantic representation. Note that Scan-RL's representation uses 6 frames, while our semantic representation only needs 2 frames.

Num. of RGB	NBV Policy	AUC ↑	FCR ↑	Acc. ↓
2	Ours w/ Scan-RL Repre.	76.31%	79.35%	0.50
	Ours (RGB-only)	**81.24%**	**87.90%**	**0.45**
6	Ours w/ Scan-RL Repre.	87.39%	95.63%	**0.38**
	Ours (RGB-only)	**87.95%**	**96.92%**	**0.38**

In addition, we preprocess the RGB images to grayscale images because we empirically find that the grayscale images featuring edges are sufficient to guide the NBV prediction and achieve slightly better performance.