Hestia: Hierarchical Next-Best-View Exploration for Systematic Intelligent Autonomous Data Collection

CHENG-YOU LU, University of Technology Sydney
ZHUOLI ZHUANG, University of Technology Sydney
TRUNG LE, University of Technology Sydney
DA XIAO, University of Technology Sydney
YU-CHENG FRED CHANG, University of Technology Sydney
THOMAS DO, University of Technology Sydney
SRINATH SRIDHAR, Brown University
CHIN-TENG LIN, University of Technology Sydney

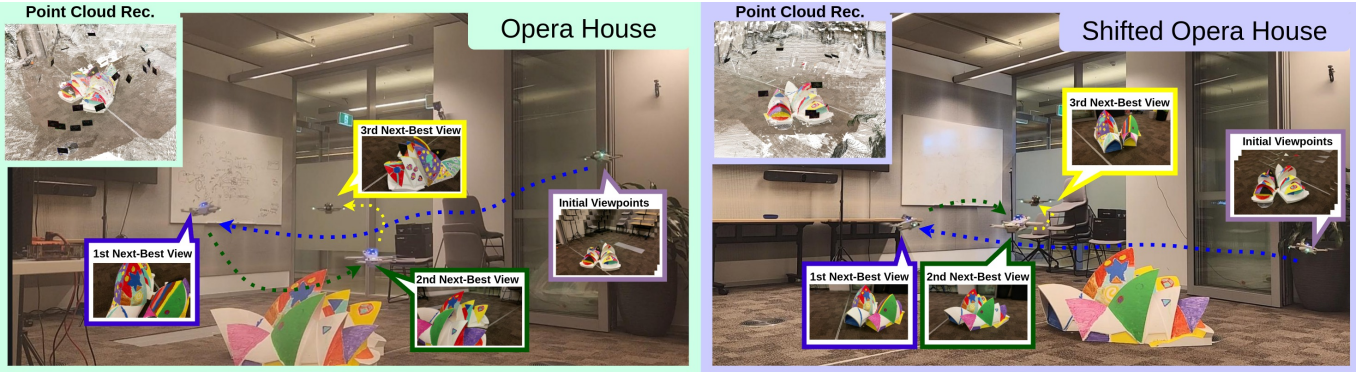

Fig. 1. Hestia is an active autonomous data collection system designed to facilitate flexible and efficient data collection. Hestia combines a next-best-view planner for viewpoint prediction with a real-world data collection system utilizing a drone equipped with an RGB camera. By integrating a pointmap estimation model [11, 68] and initializing with three viewpoints, Hestia ensures synchronization [64] between real-world and virtual environments. Hestia demonstrates robust performance, accurately predicting viewpoints in real-world scenarios even when objects are shifted. We hope that this prototype data collection system will advance the field of large-scale scene data collection using drones. *Please refer to our demonstration video in the supplementary materials for further details.*

Advances in 3D reconstruction and novel view synthesis methods have enabled efficient and photorealistic rendering. However, the data collection process for the methods is typically manual, making it time-consuming and labor-intensive. To address the challenges, this study introduces Hierarchical Next-Best-View Exploration for Systematic Intelligent Autonomous Data Collection (Hestia[1]). Hestia leverages reinforcement learning (RL) to develop a generalizable policy capable of predicting the next-best view in a 5-DoF space. Unlike prior voxel-based RL approaches, Hestia adopts the concept that a voxel is worth more than a ray, treating each voxel as a cube rather than a point for observation and reward calculation to achieve a higher coverage ratio with less data. For observation, Hestia is trained on the largest dataset for this task, ensuring robust generalization. For action, Hestia employs a hierarchical structure to separately predict the look-at point and camera position, simplifying the high-dimensional next-best-view task. For training, Hestia incorporates a greedy formulation to mitigate spurious correlations, improving training efficiency. The experimental results tested in the NVIDIA IsaacLab environment show that Hestia can achieve robust performance across datasets with objects in different positions. Hestia is integrated into the proposed real-world autonomous data collection system, demonstrating its feasibility. Unlike traditional systems, where sensors are either passive or manually controlled, Hestia's real-world deployment highlights its potential to advance active, autonomous data collection systems for capturing complex scenes. *The code and data processing script for Hestia are included in the supplementary materials and will be released after the paper is published.*

CCS Concepts: • **Computing methodologies** → **Vision for robotics**; *Markov decision processes*; Point-based models; • **Computer systems organization** → *Robotic autonomy*.

Additional Key Words and Phrases: Next-best-view planner, Reinforcement learning, Data collection system, Unmanned aerial vehicle

[1]In Greek mythology, Hestia is the goddess of the hearth, symbolizing home, foundation, and structure, representing stability and guidance in complex systems

Authors' Contact Information: Cheng-You Lu, University of Technology Sydney; Zhuoli Zhuang, University of Technology Sydney; Trung Le, University of Technology Sydney; Da Xiao, University of Technology Sydney; Yu-Cheng Fred Chang, University of Technology Sydney; Thomas Do, University of Technology Sydney; Srinath Sridhar, Brown University; Chin-Teng Lin, University of Technology Sydney.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Cheng-You Lu, Zhuoli Zhuang, Trung Le, Da Xiao, Yu-Cheng Fred Chang, Thomas Do, Srinath Sridhar, and Chin-Teng Lin. 2018. Hestia: Hierarchical Next-Best-View Exploration for Systematic Intelligent Autonomous Data Collection. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Throughout the years, multiview-based 3D scene reconstruction [1, 11, 21, 25, 45, 47, 48, 59, 60, 66, 68–71, 73] and novel view synthesis [3, 4, 8, 12, 13, 24, 40, 44, 56–58, 75, 79, 80, 82] have been critical research fields in computer vision and graphics. These methods leverage multiview information (e.g., correspondences) from images to reconstruct high-fidelity scenes or synthesize photorealistic novel views. However, collecting multiview data is often time-consuming and labor-intensive, as these methods usually depend on manually captured data. The situation becomes more challenging for generalizable approaches [3, 8, 65, 75, 79, 80, 82] that require diverse and extensive data [20, 34, 65, 76, 81]. From a data collection perspective, one way to reduce the effort is to develop machines (e.g., multiview camera rigs) [2, 29, 32, 38, 53, 72, 74] to minimize human intervention. Although these multiview data collection machines make the data collection process more scalable, their sensors are usually passive (e.g., non-movable), lacking the ability to actively explore the scene. This lack of active exploration may result in missing occluded parts during data capture. To enable active capture and alleviate the effort required for manual capture, we propose Hestia, a Hierarchical Next-Best-View Exploration for Systematic Intelligent Autonomous Data Collection. The goal of Hestia is to actively collect data from object-centric scenes with a higher coverage ratio and less amount of data, demonstrating that the proposed prototype has the potential to operate in real-world settings and to advance the development of active, intelligent data collection systems. We address this problem by introducing two main components, which are a next-best-view planner that actively determines a five-degree-of-freedom (5 DoF) viewpoint (e.g., camera position, yaw, and pitch) based on voxel face observations and a systematic intelligent, autonomous data collection system that uses a drone equipped with an RGB camera as the agent for data collection.

In recent decades, next-best-view planners have been developed for active capture, demonstrating promising potential for reducing the human workload. This is especially important when a professional human pilot is unavailable to control the robot (e.g., a drone) in time-critical tasks, such as search and rescue (SaR), where human resources are limited or urgently needed elsewhere, and in routine tasks (e.g., environmental documentation) that require frequent data collection. Traditional next-best-view planners [17, 18, 35, 42, 77] rely on heuristic rules to predict the viewpoint. Although these approaches show promising and awesome performance for specific scenarios, the same handcrafted rules or hyperparameters may not be the best setting for other scenes as discussed by [7]. Thanks to advancements in deep learning and computational capabilities, several learning-based (e.g., online learning or generalizable methods) next-best-view planners [7, 16, 22, 23, 27, 31, 49, 54, 55, 62, 78] have been developed. Among these approaches, a generalizable RL-based method [7] that utilizes an occupancy grid as the observation demonstrates effective performance in terms of coverage ratio, viewpoint flexibility, and generalization capability.

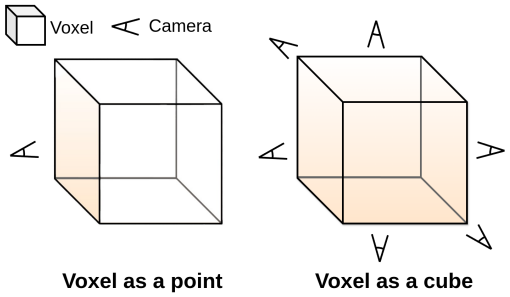

Fig. 2. **A voxel is worth more than a ray.** Unlike the RL-based generalizable method [7], Hestia treats each voxel as a cube by considering its six faces, rather than a point. This reduces the information loss inherent in point approximations, ensuring a more accurate representation of the voxel.

Inspired by the approach [7], we develop an RL-based generalizable next-best-view planner utilizing a voxel grid as the observation. While both the prior work [7] and Hestia adopt RL to learn a next-best-view policy and use voxel grids as observations, our method introduces a concept, *"A voxel is worth more than a ray"*, by incorporating the visibility of the six faces of each voxel into the observation and reward function (see Fig. 2). By explicitly representing each voxel as a cube rather than a single point, Hestia can account for the visibility of individual voxel faces, thus achieving more comprehensive data capture.

To further help the learning process, Hestia focuses on three key aspects of generalizable RL: the observation (e.g., the diversity and scale of the dataset), the action (e.g., the dimension of the action space), and the learning process (e.g., strategies for training). For observation, Hestia utilizes the largest dataset containing 30,000 shapes, processed from Objaverse [9, 10], for the next-best-view task. This ensures that Hestia encounters a more diverse range of surface geometries, rather than just cubic shapes (e.g., buildings) as shown in Fig. S1 and Fig. S2. Such diversity is potentially critical for real-world scenarios, which may involve a wide variety of shapes. For action, instead of directly predicting the 5 DoF next-best viewpoint, a task that is inherently complex and challenging, Hestia adopts a hierarchical structure to simplify the process. The approach begins by predicting the look-at point, which serves as the target of focus, followed by determining the viewpoint position based on this point. Finally, Hestia formulates the next-best-view task as a greedy optimization problem. Given an occupancy grid, the goal is to find the viewpoint that maximizes the coverage ratio. To achieve this, Hestia relies solely on the previous image, camera pose, and the given occupancy grid to predict the next-best-view, rather than depending on a sequence of images [7, 54] and camera poses [7]. Additionally, the reward discount factor γ is set to a small value to prioritize current improvements without relying on a disproportionately large goal reward. Following a greedy algorithm, Hestia can

mitigate the spurious correlations[2] (see Fig. S3) arising from false associations between the current action and a large future reward (e.g., the final goal reward). By incorporating these components, Hestia can successfully learn a generalizable next-best-view policy and reach a better coverage ratio with fewer images than previous generalizable methods [7, 23, 54].

The proposed next-best-view planner is integrated into a systematic, intelligent, autonomous data collection system, with the planner running on a ground station, to demonstrate its effectiveness in real-world scenarios. The system employs a drone as an agent, equipped with a camera mounted on a gimbal, to capture data. Using drones instead of robot arms [27] practically enables data collection in regions that are difficult for humans to access and supports operations involving large-scale outdoor objects (e.g., buildings). The contributions of this work are:

- A generalizable RL-based next-best-view planner achieves a better coverage ratio with fewer images by explicitly considering voxels as cubes rather than points.
- Help the learning process through hierarchical handling of the high-dimensional action space, training on the largest dataset to ensure diversity, and adopting a greedy algorithm to minimize spurious correlations.
- Integrate the next-best-view planner into an intelligent, autonomous data collection prototype to demonstrate its potential for active data capture under real-world scenarios.

2 Related Work

Data collection systems. Advances in 3D reconstruction [11, 45, 46, 48, 50, 59, 63, 66, 68, 69, 71] and novel view synthesis [3–5, 8, 12, 13, 24, 30, 36, 40, 43, 44, 56–58, 67, 75, 79, 80, 82, 83] techniques have enabled the rendering of high-fidelity and photo-realistic scenes. However, most approaches depended on manually captured data [14, 51, 52, 74], which is both time-consuming and labor-intensive. This limitation makes the data collection process less practical for generalizable approaches [3, 5, 8, 11, 30, 50, 65, 67, 68, 75, 79, 80, 82]. To overcome this, some methods [9, 10, 33, 34, 37, 65, 81] collected data from the internet, leveraging contributions from people worldwide. While these approaches enable the rapid collection of large datasets and benefit generalizable methods, they face potential drawbacks such as inconsistent data quality and the risk of saturation due to the inability to generate their own data. Instead, many people developed their data collection machines [2, 6, 15, 28, 29, 32, 38, 72, 74] to capture high-quality data, which is a more sustainable approach in the long term. For instance, multiview camera rigs have been used to capture face-forward scenes [2, 28, 29, 32, 74], 360° outward-facing scenes [72], and 360° inward-facing scenes [38]. Other studies [6] have combined multiple camera devices (e.g., spectacles camera and 360° camera) to capture scenes from different perspectives. Additionally, ego-centric approaches [15] developed wearable camera glasses to collect data from a first-person perspective while the user was in motion. Although these methods can collect large amounts of data using multiple cameras or passively gather data through human-operated devices, they lack the capability to actively explore the scene. In contrast, our proposed data collection system is intelligent, enabling autonomous exploration of the scene using a drone equipped with a mounted camera. While demonstrated in an indoor environment, drones as agents, unlike robotic arms, have shown the potential to capture data in hard-to-reach areas for humans and large-scale scenes (e.g., buildings).

Scene-specific next-best-view planners. Next-best-view planners have demonstrated promising results in active 3D reconstruction by predicting the optimal viewpoint for data capture based on the current state (e.g., the reconstruction progress). Traditional approaches [17, 18, 35, 42, 77] rely on hand-crafted rules to determine the next-best viewpoint. For instance, the method [77] selected the next-best viewpoint by maximizing a rating function that prioritizes smooth regions and partially overlapping patches. Similarly, the approaches [18, 42] proposed a method to identify object boundaries and collect data along the boundary between seen and unseen object surfaces, which was later extended to multi-agent settings [17], to enable more efficient data capture. Another approach [35], inspired by human behavior, scanned segmented objects sequentially, using a predefined database of similar objects to guide the selection of the next-best viewpoint. Recent advances in deep learning and increased computational power have given rise to learning-based next-best-view methods [7, 23, 27, 31, 49, 54, 55, 62, 78]. In particular, some studies [31, 62] proposed training multiple neural radiance field (NeRF) [40] models and leveraging ensemble learning to measure the disagreement among these models as an indicator of uncertainty, which was then used to select the viewpoint that maximizes this uncertainty. Other works [49, 55] incorporated Bayesian-based NeRF [39, 61] by treating radiance estimates as Gaussian distributions and using an active learning scheme to choose the viewpoint that achieves the highest information gain through variance reduction. Meanwhile, another line of approaches [27, 78] defined the next-best viewpoint as the viewpoint that maximizes the entropy of the density field along the camera rays. Although these methods have demonstrated outstanding performance in collecting data, all require sampling candidate viewpoints to determine the next-best viewpoint. In addition, because these approaches build upon NeRF, they rely on online learning and are therefore scene-specific.

Generalizable next-best-view planners. Unlike the aforementioned online-learning approaches, the generalizable methods [7, 23, 54] avoided the training process for new scenes, thereby enabling faster next-best-view selection. The prediction time is important for real-world tasks where a robot (e.g., a drone) may run out of battery within a few minutes. Among the generalizable methods, prior work [23] developed a Bayesian-based generalizable NeRF, building upon generalizable NeRF [75] and Bayesian approaches [39, 61]. Hence, the approach [23] can select the next-best viewpoint that maximizes the variance of the view for an unknown scene without requiring further training. Another line of generalizable next-best-view approaches [7, 54] utilized reinforcement learning (RL) to learn a next-best-view planner to bypass the needs of sampling candidates viewpoints. Prior work [54] proposed learning a 3-DoF next-best-view policy using a series of grayscale images as observations and

[2]Spurious correlations [19, 26] refer to certain groups contributing to model errors. In this study, spurious correlations refer to large positive future rewards assigned to suboptimal current next-best-view decisions, leading to ineffective policy learning.

the coverage ratio between the ground truth point cloud and the reconstructed point cloud as the reward. Subsequently, prior work [7] improved upon this method by incorporating occupancy grids into the observations, which provide explicit geometric information. This enhancement enabled the development of a 5-DoF next-best-view planner, achieving an outstanding coverage ratio for unknown scenes. Although the approach [7] achieved impressive results in viewpoint flexibility and coverage ratio, its simplification of treating each voxel as a point overlooked the geometry of the surface.

Different from voxel-based generalizable RL methods, Hestia integrates the concept, *"A voxel is worth more than a ray"*, into the voxel grid to enable more comprehensive data collection (see Fig. 2). Unlike prior works [7, 54], which were trained on the cubic-like House3K dataset, Hestia is trained on a larger, more diverse dataset of approximately 30,000 shapes from Objaverse [9, 10] (see Fig. S1 and Fig. S2), improving robustness in real-world scenarios. Hestia predicts the camera's look-at point first and determines the viewpoint position later, simplifying the 5-DoF action space and surpassing the 3-DoF constraints of prior works [54]. It employs a greedy optimization scheme with a small discount factor to maximize the coverage ratio and mitigate spurious correlations. Unlike methods [7, 54] that depend on sequences of images and large goal rewards, Hestia operates effectively with minimal dependencies. Integrated into a real-world data collection system, Hestia demonstrates its feasibility in practical scenarios.

3 Methods

3.1 RL Problem Definition

Our next-best-view task is to identify a 5-DoF viewpoint that maximizes the coverage ratio of an incomplete occupancy grid of the scene. The task's goal is similar to greedy methods, which always seek the locally optimal solution. We formulate the problem as a Markov Decision Process (MDP), denoted by the tuple $\{S, A, P, R, \gamma\}$. At each time step t, the agent (e.g., a drone) with RGB-D camera observes a state s_t from the set of all possible states S and chooses an action a_t from the action space A. The environment then transitions to the subsequent state s_{t+1} according to the probabilities described by P, and provides a reward r_t. The magnitude of this reward is determined by the reward function

$$R(\cdot \mid s, a) : S \times A \to r. \tag{1}$$

In reinforcement learning (RL), the main goal is to discover an optimal policy π that maximizes the expected sum of discounted rewards, given by

$$E_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \tag{2}$$

where $\gamma \in (0, 1)$ is the discount factor. We set the discount factor, γ, to 0.01 to align with the greedy-like objective and to avoid spurious correlations from large positive future rewards (see Fig. S3).

State space. The state space of Hestia is defined as

$$S = \left\{ s_t \;\middle|\; s_t = \{I_t, M_t, G_t, L_t\},\ t \in \mathbb{N} \right\} \tag{3}$$

where $I_t \in \mathbb{R}^{h \times w}$ is the grayscale image with height h and width w, and $L_t \in \mathbb{R}^3$ is the camera look-at point. The vector $M_t \in \mathbb{R}^6$ consists of $X_t \in \mathbb{R}^5$, which is the camera position, pitch, and yaw, as well as $H_t \in \mathbb{R}^1$, representing the maximum flyable height for the capture. Meanwhile, $G_t \in \mathbb{R}^{g \times g \times g \times 10}$ includes the aggregated grid information at resolution g, consisting of $O_t \in \mathbb{R}^{g \times g \times g \times 1}$ for the cumulative occupancy grid, $C_t \in \mathbb{R}^{g \times g \times g \times 3}$ for the positional encoding, and $F_t \in \{0, 1\}^{g \times g \times g \times 6}$ for the cumulative face visibility. The cumulative face visibility is updated iteratively as

$$F_t = f_t \vee F_{t-1} \tag{4}$$

where F_t represents the cumulative face visibility for all voxels up to time t, and $f_t \in \{0, 1\}^{g \times g \times g \times 6}$ denotes the face visibility for the current timestep t. To compute f_t, the depth image D_t is unprojected into a voxelized point cloud $V = \{v_i \mid i \in \mathbb{N}\}$, where v_i is the i-th voxel. Each voxel v_i is associated with a viewing direction vector $d_{v_i} \in \mathbb{R}^3$, defined as the vector pointing from the voxel center to the collision-free camera position $a_t^{'}$. The vector $\mathbf{d}_{v_i}$ is computed as

$$\mathbf{d}_{v_i} = \frac{a_t^{'} - p_{v_i}}{\|a_t^{'} - p_{v_i}\|} \tag{5}$$

where p_{v_i} is the center of voxel v_i. For each voxel v_i and its six outward-facing face normals $n_{i,j} \in \mathbb{R}^3$, the face visibility is determined and aggregated as

$$f_t(v_i, j) = \mathbb{1}\left(d_{v_i} \cdot n_{i,j} > 0\right), \quad \forall v_i \in V,\ j \in \{1, \ldots, 6\} \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function. By iterating over all voxels and their respective faces, f_t is constructed for the current timestep, and the cumulative visibility F_t is updated accordingly. Although this method cannot handle all face visibilities, the approximation enables efficient computation of face visibility. Unlike prior works [7], which consider only O_t and C_t and thereby treat voxels as points, we treat voxels as cubes to mitigate the information loss caused by approximating voxels as points (see Fig. 2). For details regarding O_t and C_t, please refer to the work [7].

Action space. The action space

$$A = \left\{ a_t \;\middle|\; a_t \in [-1, 1]^3,\ t \in \mathbb{N} \right\} \tag{7}$$

represents the set of possible 3-DoF viewpoints (e.g., camera positions) at each time step t, where each coordinate is initially bounded within $[-1, 1]$. These coordinates are subsequently normalized to the environment's scale to ensure appropriate positioning within the scene. Additionally, the camera's pitch and yaw are derived from the look-at point and the collision-free action $a_t^{'}$ converted from a_t (see Sec. 3.2).

Reward. The reward function is defined as

$$r_t = R(s_t, a_t) = r_{\text{coverage}}(s_t, a_t) + r_{\text{constraint}}(s_t, a_t) \tag{8}$$

where $r_{\text{coverage}}(s_t, a_t)$ encourages the observation of new voxel faces and is expressed as

$$r_{\text{coverage}}(s_t, a_t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{6} \left(F_t^{i,j} - F_{t-1}^{i,j}\right) \cdot M_{\text{col}}}{N \cdot 6} \cdot 0.3 \tag{9}$$

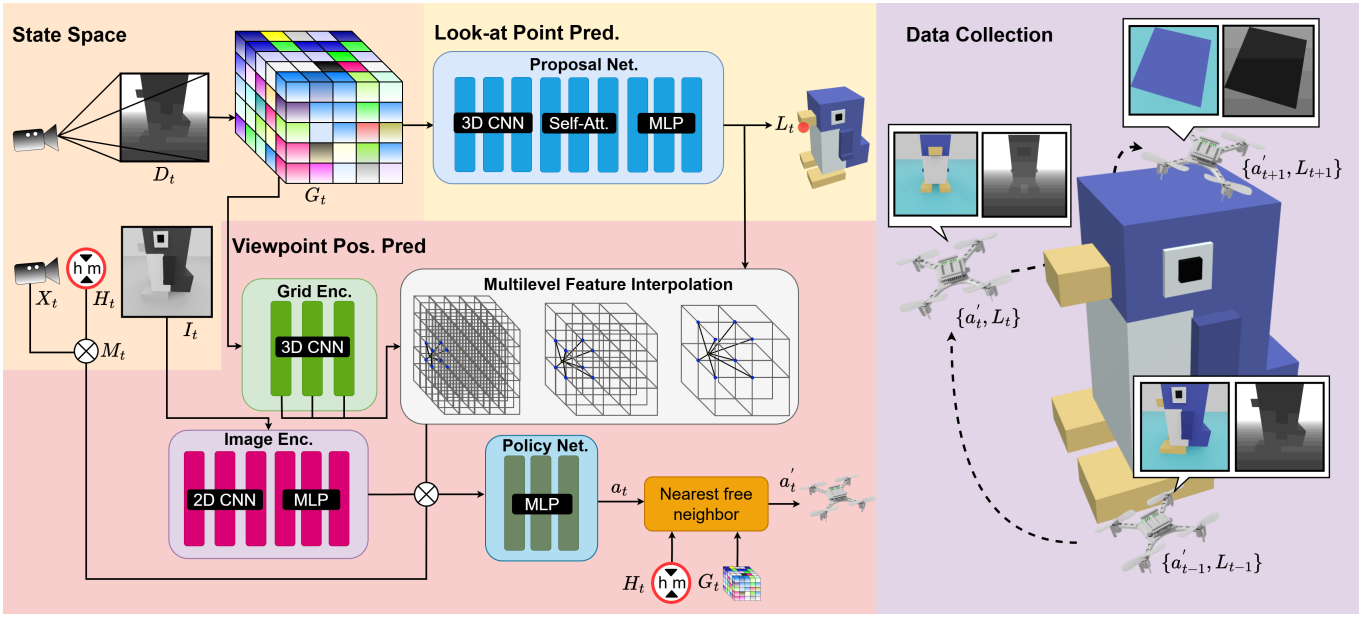

Fig. 3. **Hierarchical structure of Hestia.** Hestia first predicts the camera's look-at point L_t using a proposal neural network that takes grid information G_t processed from the depth image D_t and the camera pose as input. Next, Hestia employs a grid encoder to encode the grid information G_t and performs trilinear interpolation to extract corresponding features from the encoded grid at different layers based on L_t. These multilevel interpolated features are then concatenated with the vector information M_t which includes the camera pose X_t and the maximum flyable height, H_t as well as the encoded image features. The image features are extracted using an image encoder, which takes the grayscale image I_t as input. Finally, this combined feature representation is fed into the RL policy model to predict the camera's position a_t. Note that Hestia adopts a'_t, the nearest collision-free point to a_t, as the final camera position to ensure a collision-free viewpoint. Hence, the next-best viewpoint $\{a'_t, L_t\}$ is used for data collection.

where $F_t^{i,j}$ and $F_{t-1}^{i,j}$ represent the visibility status of the j-th face of the i-th voxel at time t and $t-1$, respectively. Here, $M_{\text{col}} \in \{0,1\}$ is a collision indicator, set to 0 in the event of a collision, thereby preventing any positive reward for invalid actions. The term $r_{\text{constraint}}(s_t, a_t)$ penalizes unsafe and invalid actions and is defined as

$$r_{\text{constraint}}(s_t, a_t) = \begin{cases} -0.01, & \text{if } r_{\text{coverage}}(s_t, a_t) = 0, \\ & \text{or } a_t[2] > H_t, \\ & \text{or } a_t \in \text{non-free voxels}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Our reward is based on the face coverage ratio rather than the point coverage ratio to ensure more comprehensive capture (see Fig. 2). Furthermore, to prevent spurious correlations, the reward design aligns with a greedy-like objective, which differs significantly from prior works [7, 54] that provide a large goal reward when the coverage ratio reaches a predefined target.

3.2 Next-Best-View Hierarchical Network

The goal of the next-best-view task is to predict a 5-DoF viewpoint for data collection. Modeling the 5-DoF viewpoint directly in the RL continuous action space is challenging due to the high-dimensional search space and complexity. To address this, Hestia introduces a hierarchical structure to simplify the problem.

Look-at point prediction. Hestia first predicts the camera's look-at point using a proposal network (see Fig. 3), which takes the grid information G_t as input. The proposal network is a shallow 3D convolutional neural network that employs a self-attention layer to expand the receptive field. The output is then passed through linear layers to decode the look-at point L_t. To model the look-at point as a probability distribution, the reparameterization trick is used, treating it as a sample from a normal distribution.

Viewpoint position prediction. To predict the remaining 3-DoF viewpoint position, the grid information is encoded into a multilevel feature grid using a shallow 3D CNN. The look-at point L_t is then used to perform trilinear interpolation on the multilevel features from the grid. These interpolated features are concatenated with the image embedding, which is extracted by an image encoder, a shallow convolutional neural network that takes the grayscale image I_t as input. Additionally, the features are concatenated with vector information M_t, which includes the camera pose X_t and the maximum flyable height H_t. The combined features are fed into the RL policy model to predict the action a_t. While the reward function helps constrain a_t to avoid collisions, an additional constraint is applied to ensure a collision-free viewpoint. Specifically, a_t is shifted to its nearest collision-free point a'_t determined using G_t and H_t. This adjusted action a'_t serves as the final viewpoint position for data capture. Thus, Hestia's next-best-view is represented as $\{a'_t, L_t\}$.

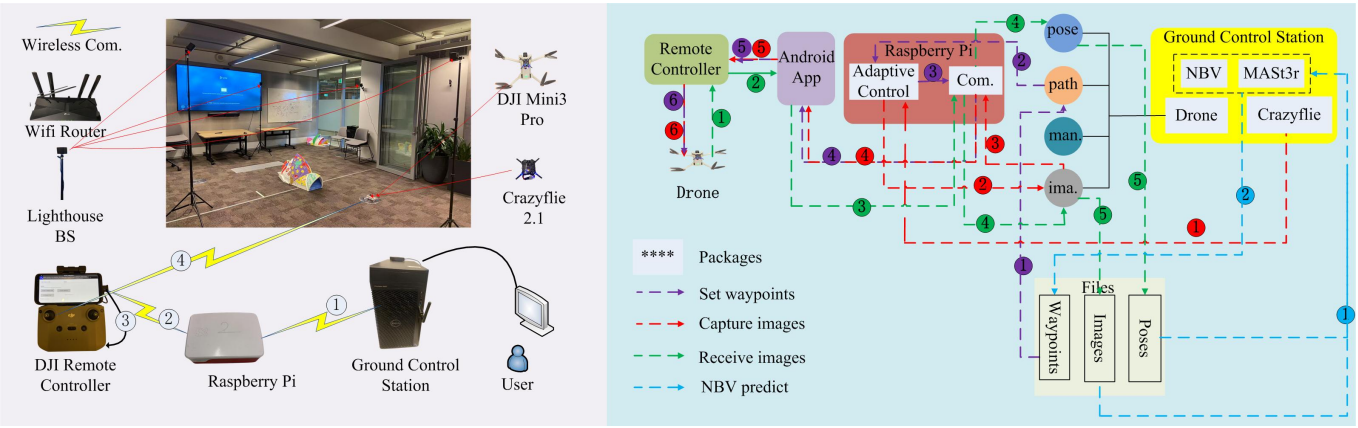

Fig. 4. **The proposed data collection system and flowchart.** Hestia goes beyond prior works by introducing a prototype of an intelligent data collection system utilizing a drone equipped with an RGB camera for data capture, HTC Lighthouse base stations and Crazyflie 2.1 for localization, and a Wifi router for wireless communication. Although tested indoors, this system highlights the potential of drones with next-best-view planners for large-scale data collection.

Training loss functions. Fully joint learning of two continuous action spaces through RL is complex and unnecessary for our task. The look-at point prediction network can instead be trained using supervised learning with a ground truth target. The ground truth look-at point L_t^{gt} is computed as the weighted average position of the ground truth uncaptured surface

$$L_t^{\text{gt}} = \frac{\sum_{v_i \in U} w_{v_i} p_{v_i}}{\sum_{v_i \in U} w_{v_i}} \tag{11}$$

where U represents the set of voxels containing ground truth uncaptured faces, and w_{v_i} is defined as the total number of ground truth uncaptured faces within voxel v_i

$$w_{v_i} = \sum_{f \in F_{v_i}^{\text{gt}}} 1 \tag{12}$$

where $F_{v_i}^{\text{gt}}$ is the set of ground truth uncaptured faces associated with voxel v_i. Thus, the loss function for the proposal network is formulated as

$$\mathcal{L}_{\text{proposal}} = \|L_t - L_t^{\text{gt}}\|^2 \tag{13}$$

The loss of the viewpoint prediction network is the same as the regular RL loss $\mathcal{L}_{\text{RL}}$ which depends on the RL method used, combined with an auxiliary loss

$$\mathcal{L}_{\text{aux}} = \|a_t - a_t^{'}\|^2 \tag{14}$$

to encourage the predicted action a_t to align with the collision-free action $a_t^{'}$. Hence, the overall loss function for Hestia is

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{RL}} + 0.5 \cdot \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{proposal}} \tag{15}$$

4 Real-world Data Collection System

To demonstrate our work, we propose a real-world system (see the left part of Fig. 4), where a drone equipped with an RGB camera moves to the next-best viewpoint predicted by Hestia to capture images of an object. The system uses four HTC Lighthouse base stations and a Crazyflie 2.1 for localization and transmits images to the ground control station via wireless communication. MASt3R [11] is integrated to convert RGB images into pointmaps (e.g., depth images), and three initial viewpoints are set for real-world and virtual-world synchronization [64]. The process (see Supp. Alg. 1) comprises four key processes, shown in the right part of Fig. 4 with four colored arrows and described in Supp. Alg. 2 to Supp. Alg. 5.

Specifically, the drone captures images at three initial viewpoints (see Supp. Alg. 1), where, for each viewpoint, process *"set waypoints"* (see Supp. Alg. 2) navigates the drone to a set of waypoints. Then, process *"capture image"* (see Supp. Alg. 3) starts the image capturing once the drone reaches the waypoint, and process *"receive image"* (see Supp. Alg. 4) transmits the captured image to the ground station. After receiving all the images captured at the initial viewpoints, process *"nbv prediction"* (see Supp. Alg. 5) predicts the next-best viewpoint based on the collected data. The image capturing and next-best viewpoint prediction loop will be repeated until sufficient images have been collected (see lines 8-13 in Supp. Alg. 1). For more details, please refer to the Supp. Sec. 4.

5 Experiments

5.1 Experimental Setup

Dataset. We use Objaverse [9, 10] as the training dataset to maximize the diversity of shapes encountered during training (see Fig. S1 and Fig. S2). For each shape, we generate the ground truth occupancy grid, voxel face visibility, and point cloud (see Supp. Sec.2 for more details). Our processed training set is two orders of magnitude larger than those used in prior works [7, 23, 54] and includes at least eighteen more categories [7, 54].

Scene setup. We use NVIDIA IsaacLab [41] to create scenes for training, running 256 scenes in parallel. Each sample is scaled to a maximum side length of 8 meters and placed within a 20m x 20m x 20m scene. During training, the object is randomly translated within the scene. For testing, the object is positioned at its original location, and the four corners for benchmarking. An RGB-D camera is mounted on the Crazyflie model, positioned 0.1 meters in front of

Table 1. **CR / CD / AUC comparison on the House3K test set with 29 collected images and an initial view per shape.** Hestia demonstrates robust performance across different object positions. Since the codebase of GenNBV [7] was not released at the time of submission, we reproduced it following their guidelines, denoted by $*$. Other baselines use the officially released model.

		House3K Test				
Train	Method	Object Center				
		(0, 0)	(4, 4)	(4, -4)	(-4, 4)	(-4, -4)
DTU	NeU-NBV [23]	0.85 / 0.26 / 0.80	0.73 / 0.52 / 0.71	0.81 / 0.30 / 0.77	0.72 / 0.53 / 0.70	0.80 / 0.32 / 0.76
House3k	ScanRL [54]	0.86 / 0.21 / 0.79	0.77 / 0.40 / 0.75	0.88 / 0.19 / 0.82	0.72 / 0.50 / 0.69	0.81 / 0.28 / 0.76
	GenNBV* [7]	0.94 / 0.12 / 0.88	0.95 / 0.10 /0.90	0.94 / 0.13 / 0.88	0.94 / 0.11 / 0.89	0.91 / 0.14 / 0.88
Objaverse	Hestia	**0.97** / **0.05** / **0.96**	**0.97** / **0.05** / **0.93**	**0.97** / **0.07** / **0.93**	**0.97** / **0.05** / **0.94**	**0.97** / **0.07** / **0.93**

Table 2. **CR / CD / AUC comparison on the Objaverse test** Hestia can achieve robust performance across different object positions.

		Objaverse Test				
Train	Method	Object Center				
		(0, 0)	(4, 4)	(4, -4)	(-4, 4)	(-4, -4)
DTU	NeU-NBV [23]	0.88 / 0.20 / 0.83	0.76 / 0.41 / 0.74	0.77 / 0.31 / 0.72	0.77 / 0.41 / 0.74	0.78 / 0.31 / 0.72
House3K	ScanRL [54]	0.87 / 0.18 / 0.82	0.80 / 0.31 / 0.78	0.82 / 0.23 / 0.77	0.77 / 0.37 / 0.76	0.80 / 0.27 / 0.76
	GenNBV* [7]	0.94 / 0.11 / 0.88	0.92 / 0.12 / 0.88	0.90 / 0.15 / 0.82	0.92 / 0.12 /0.88	0.90 / 0.14 / 0.84
Objaverse	Hestia	**0.96** / **0.07** / **0.93**	**0.96** / **0.07** / **0.92**	**0.96** / **0.06** / **0.91**	**0.96** / **0.07** / **0.93**	**0.96** / **0.08** / **0.91**

the drone. The initial viewpoint is randomly set at a collision-free point, oriented to look at the object center.

5.2 Results

Quantitative and qualitative results. We compare Hestia with three popular generalizable next-best-view methods [7, 23, 54] using the coverage ratio (CR) [7], chamfer distance (CD), and the area under the curve (AUC) [7] of the coverage ratio. Hestia demonstrates robust performance on both the House3K and Objaverse test sets (see Tab. 1 and Tab. 2), highlighting that leveraging our large-scale and diverse training set significantly enhances generalizability. Moreover, Hestia's superior performance on both test sets emphasizes that *"a voxel does deserve more than a ray"* and that avoiding spurious correlations through our greedy recipe is beneficial. Notably, Hestia exhibits consistent performance across varying object centers. This demonstrates the effectiveness of the look-at point predictor, derived from the proposed hierarchical structure (see Fig. 7). The point cloud reconstruction results using depth images collected by Hestia in Fig. 5 and Fig. 6 demonstrate the effectiveness of the proposed ideas in capturing fine-grained details. For example, Hestia successfully captures the roof soffit of the first house, the occluded door of the second house, the occluded windows of the third house, and the structure beneath the long balcony of the fourth house in Fig. 5. Similarly, Hestia can capture the fine-grained details of diverse shapes, such as the underbrim of the hat on the cactus, the bottom of the wooden log, the neck of the lamp, and the short chair. These results highlight its ability to handle challenging scenarios with occlusions and fine details.

Ablation study. We conduct an ablation study (see Tab. 3) on the three concepts of Hestia: using faces as observations, the greedy formulation with a small discount factor and no goal reward, and the hierarchical structure. For simplicity, we keep the input states consistent with Hestia, and for the non-hierarchical structure, we feed the encoded grid information into the policy model to predict 5-DoF viewpoints without feature interpolation. The results show that all these components are essential to improving the performance. In Fig. S3, we present the testing results of Hestia with and without the greedy design (corresponding to the second and third settings in Tab. 3), highlighting the potential impact of spurious correlation. Without the greedy design, some non-beneficial actions may receive positive rewards due to the large discount factor and goal rewards, leading to empty views during testing.

Table 3. **Ablations.** Integrating the proposed ideas with the Objaverse [9, 10] dataset yields top performance with fewer parameters.

	Face	Greedy	Hier.	Avg. CR ↑	Avg. CD ↓	#Pa.
Hestia				0.88	0.20	6.2M
	✓			0.90	0.17	6.2M
	✓	✓		0.95	0.09	6.2M
	✓	✓	✓	**0.96**	**0.07**	**4.9M**

Real-world demonstration. We demonstrate Hestia in the proposed real-world data collection pipeline with two scenes, opera house, and shifted opera house. In Fig. 8, the proposed data collection system, with three initial viewpoints (red rectangles) for virtual-world and real-world synchronization [64], can autonomously explore the environment and capture images (blue rectangles) through the proposed next-best-view planner, thereby reconstructing a point cloud (left part of Fig. 8).

6 Conclusion

In this work, we present Hestia, a hierarchical next-best-view exploration framework for systematic intelligent autonomous data collection. Our method addresses high-dimensional viewpoint prediction by separately predicting look-at points and camera positions. Treating voxels as cubes enables more comprehensive observation, improving coverage ratios with fewer images. The greedy design mitigates spurious correlations, ensuring efficient policy learning. Trained on the largest dataset for this task, with 30,000 diverse shapes, Hestia demonstrates robustness across varied scenarios and outperforms existing generalizable methods. Its integration into a real-world drone-based data collection system highlights its feasibility and effectiveness. We hope Hestia inspires advancements in scalable, intelligent 3D reconstruction, novel view synthesis, and autonomous exploration.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. 2011. Building rome in a day. *Commun. ACM* 54, 10 (2011), 105–112.

[2] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive Light Field Video with a Layered Mesh Representation. 39, 4 (2020), 86:1–86:15.

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19457–19467.

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *European conference on computer vision*. Springer, 333–350.

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14124–14133.

[6] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 2024. 360+x: A Panoptic Multi-modal Scene Understanding Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[7] Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. 2024. GenNBV: Generalizable Next-Best-View Policy for Active 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16436–16445.

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. 2025. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*. Springer, 370–386.

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2024. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems* 36 (2024).

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13142–13153.

[11] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. 2024. MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion. *CoRR* (2024).

[12] John Flynn, Michael Broxton, Lukas Murmann, Lucy Chai, Matthew DuVall, Clément Godard, Kathryn Heal, Srinivas Kaza, Stephen Lombardi, Xuan Luo, et al. 2024. Quark: Real-time, High-resolution, and General Neural View Synthesis. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–20.

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *CVPR*.

[14] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. 2022. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems* 35 (2022), 33768–33780.

[15] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19383–19400.

[16] Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. 2023. Macarons: Mapping and coverage anticipation with rgb online self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 940–951.

[17] Guillaume Hardouin, Julien Moras, Fabio Morbidi, Julien Marzat, and El Mustapha Mouaddib. 2020. 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)*. IEEE, 1567–1574.

[18] Guillaume Hardouin, Fabio Morbidi, Julien Moras, Julien Marzat, and El Mustapha Mouaddib. 2020. Surface-driven Next-Best-View planning for exploration of large-scale 3D environments. *IFAC-PapersOnLine* 53, 2 (2020), 15501–15507.

[19] A CAUSAL INTERPRETATION. 1971. SPURIOUS CORRELATION: A CAUSAL INTERPRETATION HERBERT A. SIMON. *Causal Models in the Social Sciences* (1971), 5.

[20] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 406–413.

[21] Hanqing Jiang, Haomin Liu, Ping Tan, Guofeng Zhang, and Hujun Bao. 2012. 3d reconstruction of dynamic scenes with multiple handheld cameras. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*. Springer, 601–615.

[22] Wen Jiang, Boshu Lei, and Kostas Daniilidis. 2023. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. *arXiv preprint arXiv:2311.17874* (2023).

[23] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. 2023. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11305–11312.

[24] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 42, 4 (2023), 139–1.

[25] Hansung Kim and Adrian Hilton. 2013. 3d scene reconstruction from multiple spherical stereo pairs. *International journal of computer vision* 104 (2013), 94–116.

[26] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. 2024. Discovering and Mitigating Visual Biases through Keyword Explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11082–11092.

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

[28] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. 2022. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems* 35 (2022), 13485–13498.

[29] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5521–5531.

[30] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2022. Efficient Neural Radiance Fields for Interactive Free-viewpoint Video. In *SIGGRAPH Asia Conference Proceedings*.

[31] Kevin Lin and Brent Yi. 2022. Active view planning for radiance fields. In *Robotics Science and Systems*.

[32] Kai-En Lin, Lei Xiao, Feng Liu, Guowei Yang, and Ravi Ramamoorthi. 2021. Deep 3d mask volume for view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1749–1758.

[33] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22160–22169.

[34] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. 2021. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14458–14467.

[35] Ligang Liu, Xi Xia, Han Sun, Qi Shen, Juzhan Xu, Bin Chen, Hui Huang, and Kai Xu. 2018. Object-aware guidance for autonomous scene reconstruction. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–12.

[36] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. 2023. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13–23.

[37] Chongshan Lu, Fukun Yin, Xin Chen, Wen Liu, Tao Chen, Gang Yu, and Jiayuan Fan. 2023. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7557–7567.

[38] Cheng-You Lu, Peisen Zhou, Angela Xing, Chandradeep Pokhariya, Arnab Dey, Ishaan Nikhil Shah, Rugved Mavidipalli, Dylan Hu, Andrew I Comport, Kefan Chen, et al. 2024. DiVa-360: The Dynamic Visual Dataset for Immersive Neural Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22466–22476.

[39] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7210–7219.

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

[41] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. 2023. Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments. *IEEE Robotics and Automation Letters* 8, 6 (2023), 3740–3747. https://doi.org/10.1109/LRA.2023.3270034

[42] Riccardo Monica and Jacopo Aleotti. 2018. Contour-based next-best view planning from point cloud segmentation of unknown objects. *Autonomous Robots* 42 (2018), 443–458.

[43] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. 2022. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3971–3980.

[44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.* 41, 4, Article 102 (July 2022), 15 pages. https://doi.org/10.1145/3528223.3530127

[45] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. 2020. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 414–431.

[46] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751* (2022).

[47] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)* 32, 6 (2013), 1–11.

[48] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. 2024. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*.

[49] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. 2022. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*. Springer, 230–246.

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

[51] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. *ICCV* (2021).

[52] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).

[53] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9054–9063.

[54] Daryl Peralta, Joel Casimiro, Aldrin Michael Nilles, Justine Aletta Aguilar, Rowel Atienza, and Rhandley Cajote. 2020. Next-Best View Policy for 3D Reconstruction. *arXiv preprint arXiv:2008.12664* (2020).

[55] Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. 2023. Neurar: Neural uncertainty for autonomous 3d reconstruction with implicit neural representations. *IEEE Robotics and Automation Letters* 8, 2 (2023), 1125–1132.

[56] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. 2024. NeRF On-the-go: Exploiting Uncertainty for Distractor-free NeRFs in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David J Fleet, and Andrea Tagliasacchi. 2024. Spotlesssplats: Ignoring distractors in 3d gaussian splatting. *arXiv preprint arXiv:2406.20055* (2024).

[58] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. 2023. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20626–20636.

[59] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. 2022. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*. Springer, 1–19.

[60] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. 2021. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 972–981.

[62] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. 2023. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9370–9376.

[63] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653* (2023).

[64] Shinji Umeyama. 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13, 04 (1991), 376–380.

[65] Matthew Wallingford, Anand Bhattad, Aditya Kusupati, Vivek Ramanujan, Matt Deitke, Aniruddha Kembhavi, Roozbeh Mottaghi, Wei-Chiu Ma, and Ali Farhadi. [n. d.]. From an Image to a Scene: Learning to Imagine the World from a Million 360° Videos. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[66] Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z Jane Wang, and Rabab Ward. 2021. Multi-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5722–5731.

[67] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*.

[68] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20697–20709.

[69] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2690–2698.

[70] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 842–857.

[71] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. 2023. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[72] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, et al. 2023. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*. 1–12.

[73] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*. 767–783.

[74] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. 2020. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5336–5345.

[75] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4578–4587.

[76] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

[77] Hongbin Zha, Ken'ichi Morooka, and Tsutomu Hasegawa. 1997. Next best viewpoint (NBV) planning for active object modeling based on a learning-by-showing approach. In *Computer Vision—ACCV'98: Third Asian Conference on Computer Vision Hong Kong, China, January 8–10, 1998 Proceedings, Volume II 3*. Springer, 185–192.

[78] Huangying Zhan, Jiyang Zheng, Yi Xu, Ian Reid, and Hamid Rezatofighi. 2022. Activermap: Radiance field for active mapping and planning. *arXiv preprint arXiv:2211.12656* (2022).
[79] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *European Conference on Computer Vision* (2024).
[80] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024. GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[81] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817* (2018).
[82] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. 2024. Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats. *arXiv preprint 2410.12781* (2024).
[83] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. 2024. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10324–10335.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

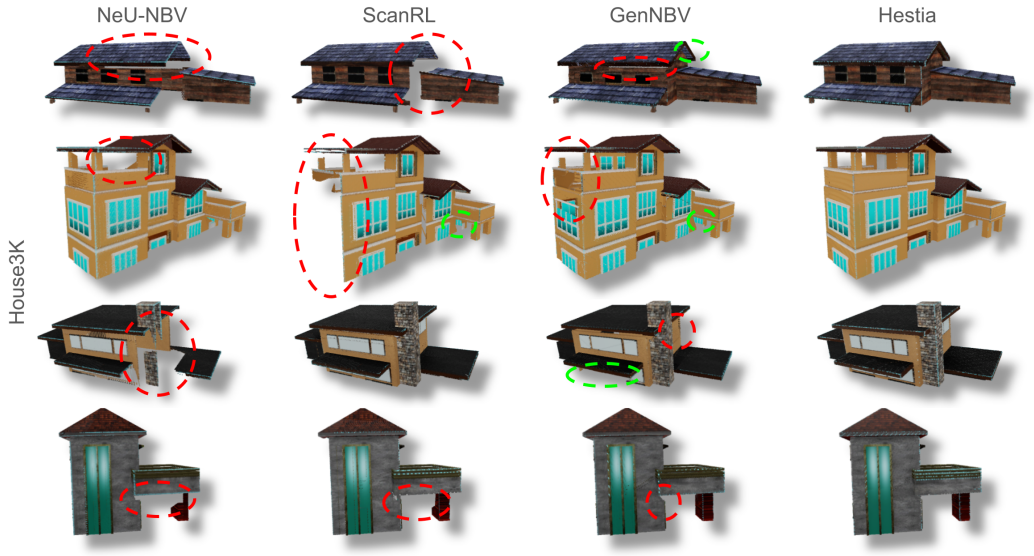

Fig. 5. **Qualitative comparison on the House3K [54].** Compared to the baselines, the point cloud reconstructed from the depth maps collected using our approach captures more fine-grained details, such as the roof soffit, particularly in self-occluded areas.

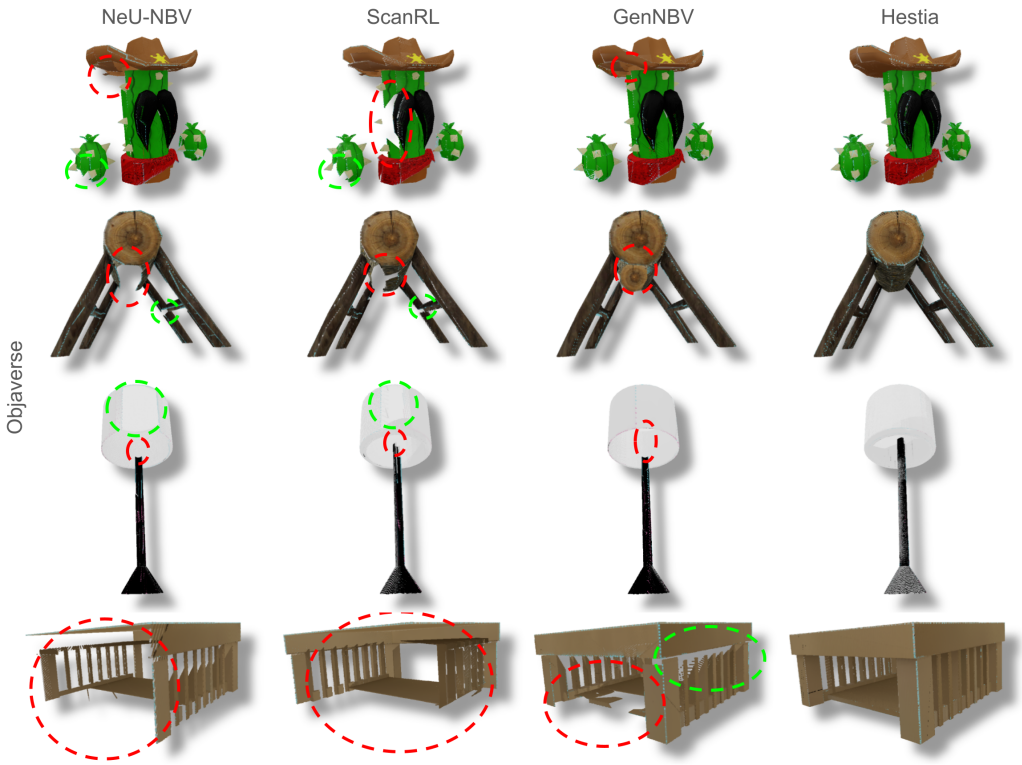

Fig. 6. **Qualitative comparison on the Objaverse [9, 10].** Compared to the baselines, the point cloud reconstructed from the depth maps collected using our approach captures finer details from bottom-up views.

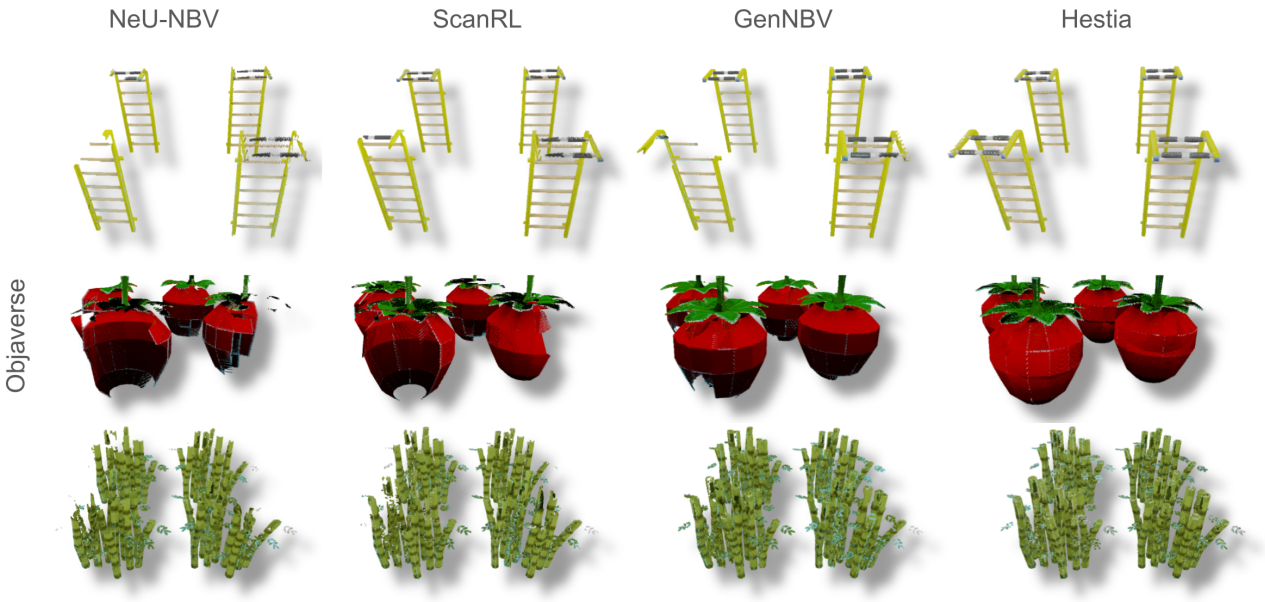

Fig. 7. **Qualitative comparison of objects at the four corners from Objaverse [9, 10].** Compared to the baselines, the point clouds reconstructed from the depth maps collected using our approach are more robust and consistent across different object centers.

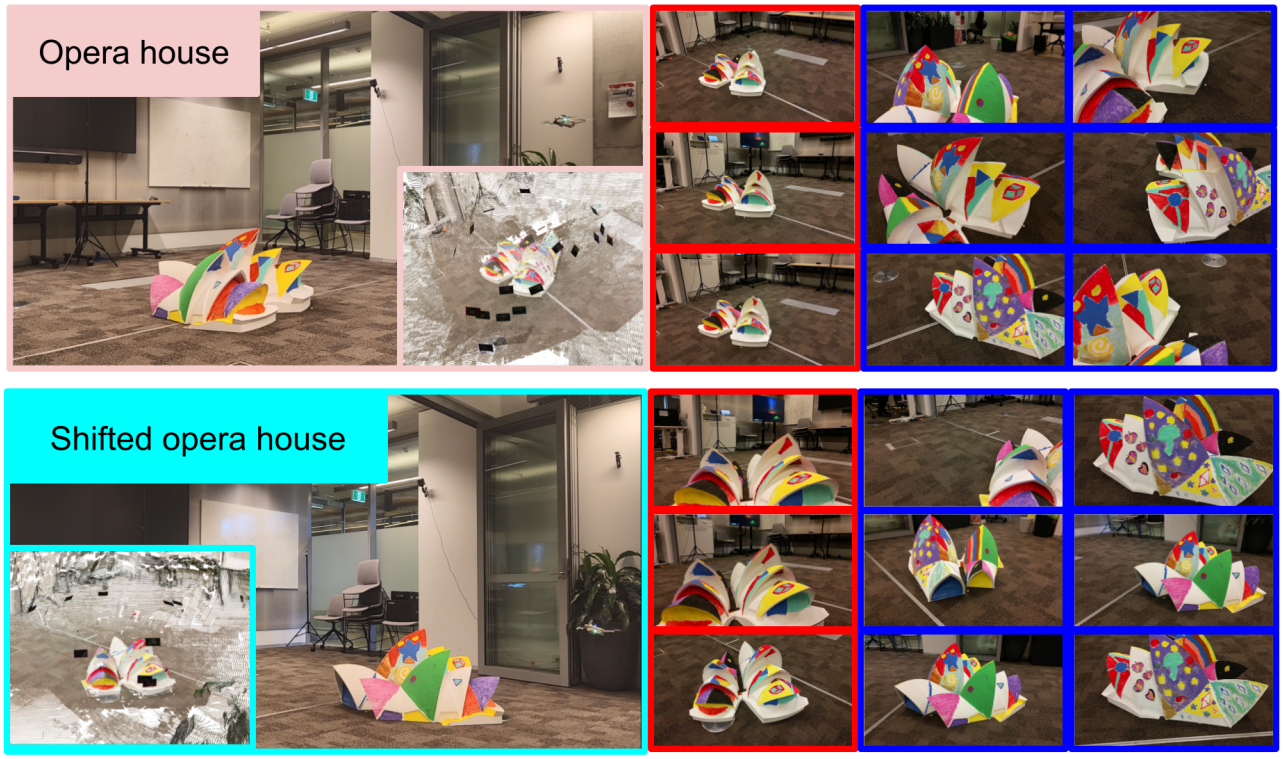

Fig. 8. **Real-world demonstration.** Hestia can operate in real-world scenarios with three initial viewpoints (red rectangles) and predict the next-best viewpoints for capture (blue rectangles for the first six views), even when the depth camera is unavailable. Point cloud reconstruction results are shown on the left, with black rectangles representing camera poses.