Exploiting Priors from 3D Diffusion Models for RGB-Based One-Shot View Planning

Sicong Pan* Liren Jin* Xuying Huang Cyrill Stachniss Marija Popović Maren Bennewitz

***Abstract*— Object reconstruction is relevant for many autonomous robotic tasks that require interaction with the environment. A key challenge in such scenarios is planning view configurations to collect informative measurements for reconstructing an initially unknown object. One-shot view planning enables efficient data collection by predicting view configurations and planning the globally shortest path connecting all views at once. However, prior knowledge about the object is required to conduct one-shot view planning. In this work, we propose a novel one-shot view planning approach that utilizes the powerful 3D generation capabilities of diffusion models as priors. By incorporating such geometric priors into our pipeline, we achieve effective one-shot view planning starting with only a single RGB image of the object to be reconstructed. Our planning experiments in simulation and real-world setups indicate that our approach balances well between object reconstruction quality and movement cost.**

I. Introduction

Many autonomous robotic applications require 3D models of objects to perform downstream tasks, e.g., pose estimation [35], object manipulation [3], and detection [36]. When deployed in initially unknown environments, a robot often needs to reconstruct the objects before interacting with them. During this procedure, a challenge is planning a view sequence to acquire the most informative measurements to be integrated into the reconstruction system while minimizing the robot's travel distance or operation time.

Without any prior knowledge about the environment, a common strategy is to plan the next-best-view (NBV) iteratively based on the current reconstruction state [7, 18, 19, 23, 29, 33, 37]. However, NBV planning only generates a local path to the subsequent view and cannot effectively distribute the mission time or movement budget, resulting in suboptimal view planning performance. An alternative line of work considers one-shot view planning [6, 26, 28]. Given initial measurements of an object to be reconstructed, one-shot view planning predicts a set of views at once and computes the globally shortest path connecting them. A robot's sensor then follows the planned path to collect measurements, which are used for object reconstruction after the data acquisition is completed. By decoupling data collection and object reconstruction, these approaches do not rely on iterative map updates for object-specific view planning online during a mission. To perform one-shot view planning, prior knowledge about the object is required. Previous works consider planning priors based on multi-view images or partial point cloud observations. However, they either only handle a fixed view configuration [28] or rely on depth sensors [6, 26].

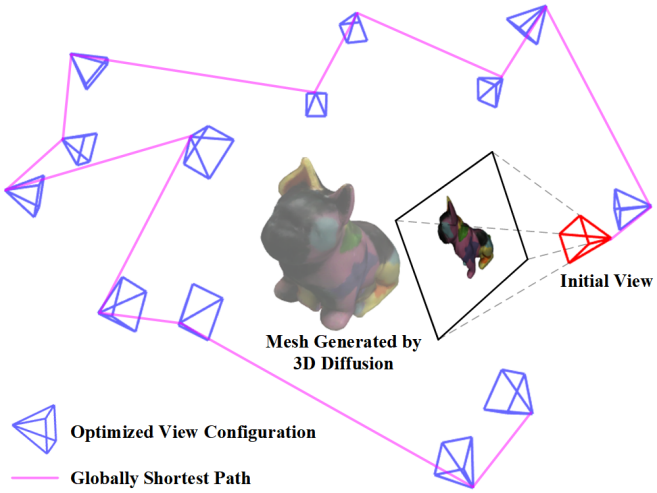

Fig. 1: An example of our RGB-based one-shot view planning by exploiting priors from 3D diffusion models. Our goal is to plan a set of views (blue) at once to collect informative RGB images for object reconstruction. The key component in our approach is a 3D diffusion model generating the corresponding 3D mesh of a single RGB image from the initial camera view (red). By leveraging the mesh as geometric priors, our approach produces view configurations specifically associated with the target object and calculates the globally shortest path. In particular, we plan denser views to observe more geometrically complex parts (front part of the object in the example) to improve the reconstruction quality.

To address these aforementioned limitations, we propose integrating geometric priors from 3D diffusion models into one-shot view planning. Recently, 3D diffusion models emerged as a powerful tool for generating 3D content based on text prompts or a single image. By training on large datasets, 3D diffusion models learn prior knowledge about objects commonly seen in real life [13, 16, 17]. Humans similarly exploit prior knowledge to hallucinate 3D models of an object based on semantics and appearance information contained in RGB observations. However, recovering a 3D representation from a single RGB image is inherently an ill-posed problem and corresponds to multiple plausible solutions. As a result, models generated by 3D diffusion models do not reflect the exact representation of the object to be reconstructed. This prohibits their direct application as a method for accurate 3D representation, as required in robotics tasks. Incorporating the capabilities of 3D diffusion models to provide geometric priors in robotics remains an

*These authors contributed equally to this work.

All authors are with the University of Bonn, Germany. Cyrill Stachniss and Maren Bennewitz are additionally with the Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany. This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 459376902 – AID4Crops and under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob. Corresponding: span@uni-bonn.de.

unexplored area.

The main contribution of this work is a novel RGB-based one-shot view planning approach that exploits the geometric priors from 3D diffusion models. Our approach enables view planning with an object-specific view configuration for object reconstruction as shown in Fig. 1. A key component of our pipeline is a 3D diffusion model that outputs a 3D mesh of the object given one RGB image as input. This generated mesh is a proxy to the inaccessible ground truth 3D model and serves as the basis for our one-shot view planning. Given the generated 3D mesh, we convert the one-shot view planning into a customized set covering optimization problem to calculate the minimum set of views that densely covers the mesh, which we solve using linear programming. Our approach places the object-specific views and follows the globally shortest path for collecting informative RGB images around the object. After the data collection, we train a Neural Radiance Field (NeRF) using all collected images to acquire the object's 3D representation.

To the best of our knowledge, our approach is the first to leverage 3D diffusion models for view planning. We make the following claims: (i) we exploit the powerful 3D diffusion models to enable our one-shot view planning starting with only one RGB image as input; (ii) we design the one-shot view planning as a customized set covering optimization problem, yielding view configurations suitable for RGB-based object reconstruction using NeRFs. We conduct extensive experiments on publicly available object datasets and in real world scenarios, demonstrating the applicability and generalization ability of our approach. Our one-shot view planning allows for object-specific view placement to account for varying object geometries, achieving a better trade-off between movement cost and reconstruction quality compared to baselines. To support reproducibility and future research, our implementation is open-sourced at: `https://github.com/psc0628/DM-OSVP`

II. Related Work

In this section, we introduce relevant works on view planning for object reconstruction and diffusion models for 3D generation.

A. View Planning for Object Reconstruction

Object reconstruction is essential in many robotic applications. One important capability in this scenario is to actively reconstruct the object using a robot sensor. Without any prior knowledge, a common approach is to plan the NBV iteratively based on the current reconstruction state, thus maximizing the information of the object in a greedy manner. Isler et al. [7] propose selecting the NBV by calculating the information gain based on visibility and the likelihood of observing new parts of the object to be reconstructed. Similarly, Pan et al. [24, 25] weight the 3D space based on visibility and distance to observe surfaces and then employ coverage optimization for NBV planning. In addition, Menon et al. [19] introduce a shape completion method based on partially observed objects and conduct NBV planning to cover the estimated missing surfaces. PC-NBV [37] trains a neural network to predict the utility of candidate views given partial point cloud observations. In the context of NBV planning for RGB-based object reconstruction, Jin et al. [8] integrate uncertainty estimation into image-based neural rendering to guide NBV selection in a mapless way. Lin et al. [11] and Sünderhauf et al. [33] train an ensemble of NeRF models, utilizing the ensemble's variance to measure uncertainty for NBV planning.

While showing promising object reconstruction results, NBV planning often relies on computationally intensive online map updates and its greedy nature leads to inefficient paths. To address these limitations, recent works propose one-shot view planning paradigm. Given an initial measurement, one-shot view planning predicts all required views at once and calculates the globally shortest path connecting them, resulting in reduced movement costs. The pioneering work SCVP [26] trains a neural network in a supervised way to directly predict the global view configuration given initial point cloud observations. To generate training labels, the authors solve the set covering problem to obtain a view configuration fully covering the ground truth 3D models. Hu et al. [6] further reduces the required views by incorporating a point cloud-based implicit surface reconstruction method to complete missing surfaces before conducting one-shot view planning. In the domain of RGB-based object reconstruction, Pan et al. [28] propose a view prediction network to predict the number of views to reconstruct an object using NeRFs required to reach its performance upper bound. However, due to the lack of geometric representations during the view planning stage, this work only considers distributing the views following a fixed pattern, without adapting view configurations to account for varying object geometries.

Our work shares the same idea of using one-shot view planning to reconstruct an unknown object. Different from previous works that rely on depth sensors [6, 26] or fixed view configurations [28], our novel approach only requires RGB inputs and plans view configurations specifically associated with the objects, leading to better object reconstruction performance while reducing movement costs.

B. Diffusion Models for 3D Generation

Diffusion models are state-of-the-art generative models for producing plausible high-quality images. Starting from random Gaussian noises, diffusion models learn to subsequently denoise the input to finally recover the true images [9, 31]. By training on large datasets, diffusion models acquire powerful prior knowledge and show their capabilities in the domain of 2D image generation.

Inspired by the advances of diffusion models, recent works investigate using diffusion models for 3D content generation. Given a text prompt describing a desired scene, DreamFusion [30], ProlificDreamer [34], and MVDream [32] optimize a differentiable 3D representation, e.g., NeRF, from scratch and leverage neural rendering to generate 2D images at different viewpoints. These rendered images are then fed into 2D diffusion models to calculate the similarity to

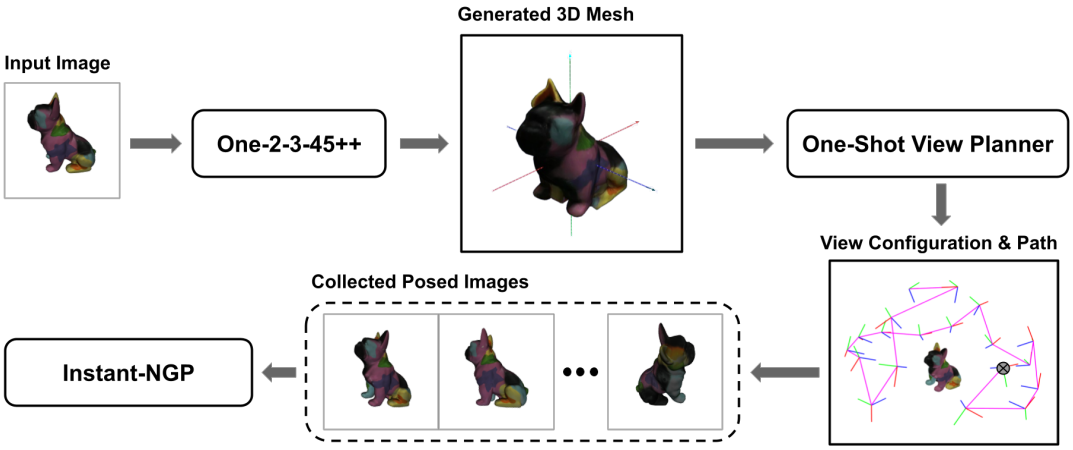

Fig. 2: Overview of our proposed RGB-based one-shot view planning pipeline. Given a single RGB image of the object to be reconstructed, we leverage a 3D diffusion model, One-2-3-45++ [12], to generate a 3D mesh. This mesh serves as a proxy to the ground truth geometry and is the basis for our view planning. Based on this prior, we construct the one-shot view planning task as a customized set covering optimization problem and solve it to obtain a minimum set of views required to densely cover the mesh surfaces. The RGB camera starts at the initial view (shown as $\otimes$) and follows the generated globally shortest path to collect RGB images, which we use to train a NeRF in Instant-NGP [21] after the data acquisition is completed.

the priors learned by the diffusion model, which guide the 3D shape optimization process. While showing impressive results, these methods suffer from prolonged rendering and optimization times, limiting their robotic applications.

Another line of work investigates fine-tuning pretrained 2D diffusion models for multi-view synthesis from single image inputs [14, 15]. The follow-up work One-2-3-45 [13] produces 3D meshes using images generated from the multi-view diffusion models. However, its performance is limited by the inconsistency between multi-view images. Recent 3D diffusion model One-2-3-45++ [12] mitigates the problem of inconsistencies by conditioning the multi-view image generation on each other. The generated multi-view consistent images are exploited as the guidance for 3D diffusion to directly produce high-quality meshes in a short time, i.e., within $60\,\mathrm{s}$. In this work, we utilize geometric priors from 3D diffusion models to enable RGB-based one-shot view planning for object reconstruction.

III. Our Approach

We propose a novel RGB-based one-shot view planning method for unknown object reconstruction. An overview of our approach is shown in Fig. 2. Given a single RGB measurement of the object, we leverage a 3D diffusion model to generate its corresponding mesh. Based on rich prior information contained in the generated mesh, we formulate one-shot view planning as a set covering optimization problem, which we solve with linear programming to acquire the minimum set of views densely covering mesh surfaces. We calculate the globally shortest path connecting all views for data collection using a robot's RGB camera. After data collection, we train a NeRF model using all collected RGB images to generate a 3D representation of the object.

A. Geometric Priors from 3D Diffusion Model

A key component of our approach is a 3D diffusion model for predicting the corresponding mesh given only one RGB image as an initial observation. Specifically, we use the state-of-the-art 3D diffusion model One-2-3-45++ [12] for generating plausible meshes due to its accurate mesh generation and efficient inference compared to other 3D diffusion models [30, 32, 34]. One-2-3-45++ model is trained on Objaverse [2], a large 3D model dataset, to learn the prior knowledge of varying geometries of commonly seen objects and shows good generalization ability on other object datasets. Leveraging this powerful tool, we use the generated meshes as geometric priors for one-shot view planning introduced next.

B. One-Shot View Planning as Set Covering Optimization

One-shot view planning can be treated as a conventional set covering optimization problem. Since solving this optimization problem necessitates an explicit 3D representation of the object to be reconstructed, previous works [6, 26] rely on depth sensors to acquire initial 3D models of the object. Instead, by incorporating the geometric priors of 3D diffusion models into our planning pipeline, our approach solves the one-shot view planning problem in an RGB camera setup.

To facilitate the efficiency of set covering optimization, sparse surface representations are desired. To this end, we first sample a set of surface points from the mesh produced by the 3D diffusion model and subsequently voxelize them using OctoMap [5] to get a sparse surface point set $\mathcal{P}_{surf}$, with surface point $p_i \in \mathcal{P}_{surf}$. We denote v as a candidate view within a discrete candidate view space $\mathcal{V} \subset \mathbb{R}^3 \times SO(3)$ and $\mathcal{P}_v$ as the set of surface points observable from this view. Each set $\mathcal{P}_v$ is determined via the ray-casting process implemented in OctoMap. We define an indicator function $I(p, v)$ to represent whether a surface point p is observable from view v:

$$I(p, v) = \begin{cases} 1 & \text{if } p \in \mathcal{P}_v \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

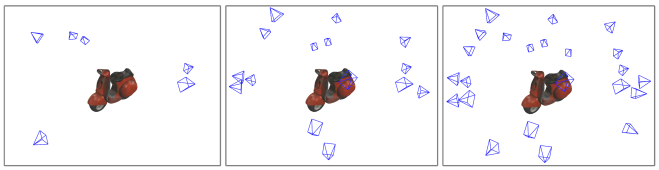

Fig. 3: Illustration of the impact of multi-view constraints. α denotes the minimum number of views required to observe each surface point. Larger α values lead to optimization solutions with more views densely covering the surfaces.

Given $\mathcal{P}_{surf}$ and each $\mathcal{P}_v$, the conventional set covering optimization problem aims to find the minimum set of views required for completely covering the surface points. For instance, consider $\mathcal{P}_{surf} = \{p_1, p_2, p_3\}$, $\mathcal{P}_{v_1} = \{p_1, p_2\}$, $P_{v_2} = \{p_2, p_3\}$, and $\mathcal{P}_{v_3} = \{p_1, p_3\}$. The union of these three sets equals the entire surface set, i.e., $\bigcup_v \mathcal{P}_v = \mathcal{P}_{surf}$. However, we can cover all surface points with only two sets, $\mathcal{P}_{v_1}$ and $\mathcal{P}_{v_2}$. Vanilla set covering optimization problem requires that each surface point should be covered by at least one view. This definition aligns well with object reconstruction employing depth-sensing modalities [6, 26, 27], as surfaces can be recovered by direct depth fusion when provided with a corresponding point cloud observation. However, for RGB-based object reconstruction using NeRFs, map representation learning is achieved by minimizing the photometric loss when reprojecting hypothetical surface points back to 2D image planes, which requires that a surface point should be observed from different perspectives to recover its true 3D representation. This implies that planned views covering all surface points of the generated mesh once are not sufficient for object reconstruction using NeRFs.

To this end, we customize the set covering optimization problem for RGB-based object reconstruction using NeRFs. Rather than requiring each surface point to be observed by at least one view, we propose multi-view constraints to enforce that a given surface point should be covered by a minimum number $\alpha \in \mathbb{N}^+$ of views to account for multi-view learning in NeRFs. Larger α values require denser surface coverage in our optimization problem, resulting in solutions with more views required as shown in Fig. 3. Note that when $\alpha \geq 2$, we exclude points that are visible from fewer than α views. This mechanism ensures the optimization problem has a feasible solution. However, our multi-view covering setup may contain multiple feasible solutions since most of the surface points can be observed from a large range of view perspectives. Some of them lead to views clustered closely together in Euclidean space. Fig. 4(a) illustrates an instance of spatially clustered views for covering the Motorbike object. These spatially clustered views exhibit similarity in the collected images, thus leading to redundant information about the object.

To alleviate this issue, we introduce a parameter $\beta \in \mathbb{R}^{\geq 0}$ for additional distance constraints to avoid selecting spatially clustered views. We denote $d_v^{v'}$ as the Euclidean distance between views v and v', while d_v^{min} is the Euclidean distance from view v to its nearest neighboring view. We prevent other views within a specific distance $\beta\, d_v^{min}$ of the view

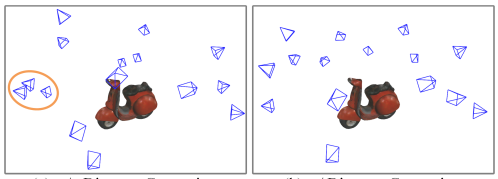

Fig. 4: Illustration of the impact of distance constraints: (a) spatially clustered views (the orange circle showcases an example of clustered views); (b) spatially more uniform views. Both view configurations are feasible solutions. By incorporating distance constraints, we express the preference for spatially uniform distribution to avoid redundant information in clustered views.

v from being selected again in the solution. A larger β leads to more spatially uniform views, while an excessively large value can render the problem infeasible. For our view planning, we try to find the maximum β value that still yields an optimization solution. Given that different objects exhibit diverse geometries, their respective maximum β values also vary. Therefore, we run optimization iteratively to find the maximum β for a specific object in an automatic manner.

Taking all these conditions into account, we formulate our set covering optimization problem as a constrained integer linear programming problem defined as follows:

$$
\begin{aligned}
\min: \quad & \sum_{v \in \mathcal{V}} x_v \,, \\
\text{s.t.}: \quad & (a) \quad x_v \in \{0,1\} && \forall v \in \mathcal{V} \\
& (b) \quad \sum_{v \in \mathcal{V}} I(p,v)\, x_v \geq \alpha && \forall p \in \mathcal{P}_{surf} \\
& (c) \quad x_v + x_{v'} \leq 1 && \forall d_v^{v'} \leq \beta\, d_v^{min},
\end{aligned}
\tag{2}
$$

where the objective function $\sum_{v \in \mathcal{V}} x_v$ is designed to minimize the total number of selected views, while subject to three constraints: (a) x_v is a binary variable representing whether a view v is included in the set of selected views or not; (b) each surface point $p \in \mathcal{P}_{surf}$ must be observed by a minimum of α selected views; and (c) if a view v is selected, any neighboring view v', whose distance $d_v^{v'}$ is smaller than $\beta\, d_v^{min}$, must not be selected.

We employ the Gurobi optimizer, a linear programming solver [4], to compute the solution for the problem. We present an instance solution in Fig. 4(b) showcasing the optimized minimum set of views required for densely covering the Motorbike object surface with $\alpha = 3$ and distance constraints.

C. Path Generation and Object Reconstruction

By planning all required views before data collection, the one-shot view planning paradigm shows a major advantage in reduced movement costs. Given the optimized set of views introduced above, we plan the globally shortest path connecting all views by solving the shortest Hamiltonian path problem on a graph, which is similar to the traveling salesman problem [22]. The robot's RGB camera follows the global path to acquire RGB measurements at planned views. We follow the point-to-point local path planning method [27] to avoid collisions with the object.

α	Planned Views	PSNR ↑	SSIM ↑	Movement Cost (m) ↓	Inference Time (s) ↓
1	6.8 ± 1.5	30.167 ± 0.810	0.9365 ± 0.0121	1.754 ± 0.258	140.4 ± 26.9
2	12.8 ± 1.7	31.436 ± 0.622	0.9530 ± 0.0049	2.629 ± 0.224	145.9 ± 29.3
3	17.8 ± 2.4	31.853 ± 0.615	0.9599 ± 0.0038	2.998 ± 0.225	147.9 ± 31.8
4	22.5 ± 3.8	31.995 ± 0.684	0.9633 ± 0.0035	3.214 ± 0.372	148.2 ± 33.1
5	28.7 ± 3.8	32.120 ± 0.786	0.9663 ± 0.0034	3.725 ± 0.312	150.0 ± 40.6
6	34.1 ± 5.1	*32.243 ± 0.779	*0.9684 ± 0.0042	4.093 ± 0.441	147.6 ± 34.1
7	38.8 ± 3.8	†32.248 ± 0.807	†0.9694 ± 0.0041	4.190 ± 0.247	147.3 ± 38.2

TABLE I: Analysis on multi-view constraints. α denotes the minimum number of views required to observe each surface point. Planned views indicate the number of optimized views under different α values. PSNR and SSIM are averaged over 100 novel views. Each value reports the average mean and standard deviation on 10 test objects. The star symbol ($\star$) indicates statistically significant results for $\alpha = 6$ compared to $\alpha = 5$ based on the paired *t*-test with a *p*-value of 0.05. Conversely, the dagger symbol ($\dagger$) indicates non-significant results for $\alpha = 7$ compared to $\alpha = 6$ based on the paired *t*-test with a *p*-value of 0.05. Results show that our optimizer plans more views with increasing α values and achieves peak performance at the $\alpha = 6$. It is worth mentioning that increasing α from 1 to 2 leads to the highest performance gain, indicating that our formulation of set covering benefits NeRF-based reconstruction.

Fig. 5: Ten test objects used in our simulation experiments.

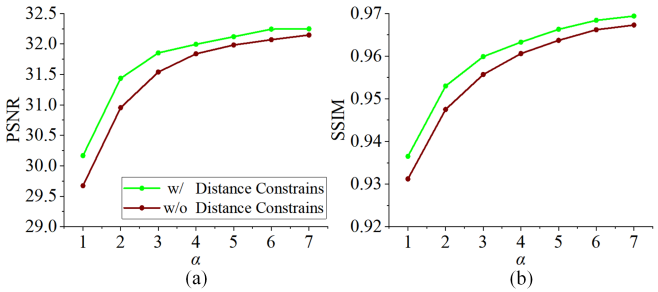

Fig. 6: Ablation study on distance constraints. PSNR and SSIM averaged over 100 novel views. Each value is reported as the averaged mean on 10 test objects. We observed statistically significant results for our method when compared to the version without distance constraints across all α values, as determined through paired *t*-tests with a *p*-value of 0.05. This suggests that the set covering optimization with the distance constraints finds better view configurations, leading to superior NeRF training results.

After data collection, we use NeRFs to acquire the final 3D representation of the object. Specifically, we adopt Instant-NGP [21] to train our NeRF, due to its efficient training performance and common usage in baseline approaches [11, 28, 33].

IV. Experimental Results

A. Experimental Setup

In our simulation experiments, we consider an object-centric hemispherical view space with 144 uniformly distributed view candidates for view planning [28]. We set the view space radius to 0.3 m. We test our approach on 10 geometrically complex 3D object models from the HomebrewedDB dataset [10]. The test objects are shown in Fig. 5. We normalize all objects to fit into a bounding sphere with a radius of 0.1 m. All RGB measurements are at 640 px × 480 px resolution. We adopt a grid size of $50 \times 50 \times 50$ in OctoMap for voxelizing the mesh surface points. The set covering optimization for view planning runs on an Intel i7-12700H CPU, while NeRFs training is conducted on an NVIDIA RTX3060 laptop GPU.

To evaluate NeRF reconstruction quality, we report the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [20] on 100 uniformly distributed novel views [28]. Additionally, we evaluate reconstruction efficiency by inference time for view planning and accumulated movement cost for data collection in Euclidean distance.

B. Analysis on Multi-View Constraints

In this section, we explore the influence of multi-view constraints introduced in Sec. III-B. We test our methods across varying α values from 1 to 7, as detailed in TABLE I. The outcomes reveal that: (1) with increasing α values, our optimizer outputs on average more views for covering the mesh surfaces; (2) both PSNR and SSIM metrics exhibit a consistent improvement with increasing α. Specifically, we achieve the highest performance gain by changing $\alpha = 1$ to $\alpha = 2$, justifying our modification of the set covering optimization problem to account for RGB-based object reconstruction using NeRFs; (3) our method reaches its peak performance at the α value of 6, while increasing α to a higher value does not yield a statistically significant performance improvement.

C. Ablation Study on Distance Constraints

In this ablation study, we investigate the impact of the distance constraints introduced in Sec. III-B. To prevent the optimizer from finding a view configuration that leads to clustered views, we introduce the parameter β as the distance constraints into our optimization formulation. We adopt binary search in our implementation to find out the object-specific maximum β that still yields a feasible optimization solution. The search step is set to 0.1 for all experiments.

We evaluate the influence of our distance constraints by performing an ablation study over different α values. Fig. 6 shows the differences between optimization with and without the proposed constraints. In all circumstances, optimization without considering the distance constraints ($\beta = 0$) outputs clustered views with redundant information about the object,

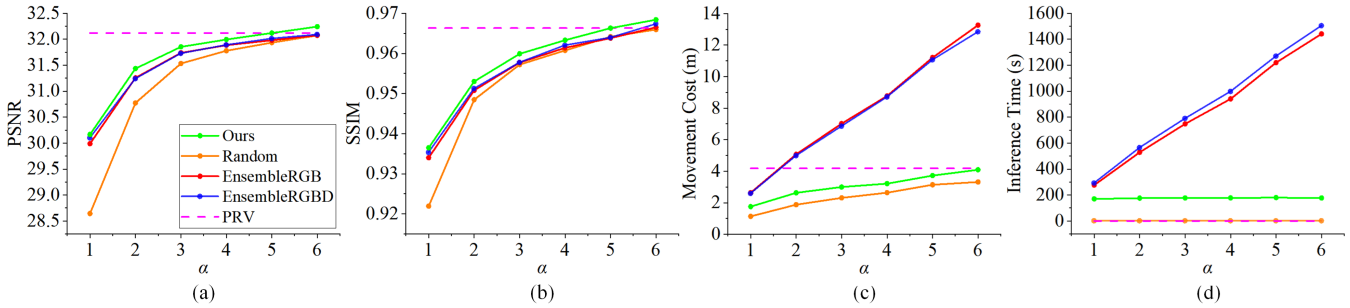

Fig. 7: Comparison to baselines on view planning performance under different α values corresponding to the number of optimized views. PSNR and SSIM are averaged over 100 novel views. Each value reports the mean on 10 test objects. PRV is not associated with α values and is represented by a dashed line. As can be seen, (1) our method achieves higher PSNR/SSIM values against *Random* and NBV methods, indicating that leveraging geometric priors from diffusion models leads to more informative views; (2) compared to PRV using fixed view configuration, our object-specific view configuration is more suitable for view planning, achieving either a lower movement cost with an on-par performance ($\alpha = 5$) or a higher performance with a slightly lower movement cost of 0.09 m ($\alpha = 6$).

leading to inferior NeRF training performance in terms of PSNR and SSIM. This justifies our design choice of introducing the distance constraints to find better view configurations.

D. Evaluation of View Planning for Object Reconstruction

Baselines. We compare our novel one-shot view planning with the following baselines:

- *Random* selects a certain number of views randomly and subsequently plans a global path to connect them.
- *EnsembleRGB* [11] leverages RGB variance of the NeRF ensemble as uncertainty quantification to plan the NBV that maximize the information gain.
- *EnsembleRGBD* [33] extends EnsembleRGB by incorporating a density-aware epistemic uncertainty computed on ray termination probabilities in unobserved object areas.
- *PRV* [28] uses a network to predict the required number of views that achieves the peak performance of NeRF training. A fixed hemispherical view configuration is then generated according to the predicted number of views.

For a fair comparison, we use Instant-NGP [21] with the same configuration for the training and testing in all experiments. Therefore, the performance differs purely as a consequence of collected RGB images using different planning strategies. As depicted in TABLE I, varying α values result in different numbers of planned views. Therefore, to comprehensively assess the performance of our planner, we evaluate all baselines using an equivalent number of views corresponding to each α value in our approach (excluding PRV, which predicts its own required number of views).

Comparison to Random Selection. As shown in Fig. 7, our RGB-based one-shot view planning approach surpasses the one-shot *Random* baseline across all α values in terms of PSNR and SSIM. This is because the heuristic *Random* method does not utilize any available information about the objects, in contrast to our approach. The *Random* method exhibits a slightly lower movement cost. We believe that this occurs since it can produce spatially clustered views, yielding poorer reconstruction quality. These findings confirm that leveraging powerful geometric priors from 3D diffusion models significantly benefits one-shot view planning for RGB-based object reconstruction.

Comparison to NBV Methods. Compared to two NBV baselines, our method achieves higher PSNR and SSIM values across all α values with much less movement costs and inference time, as shown in Fig. 7. Specifically, our method excels under various resource constraints, e.g., different planned views according to different α values. This implies that using diffusion models for priors leads to more informative views for unknown object reconstruction compared to NBV methods considering the ensemble's variance for uncertainty measurements. We attribute the significant reductions in movement cost and inference time to global path planning and the one-shot non-iterative paradigm, which avoids iterative map updates and uncertainty computation.

Comparison to PRV. Since the PRV method obtains the number of views by predicting the upper limits of NeRF representations, it is not associated with α values and is represented by a dashed line in Fig. 7. The results indicate that the proposed RGB-based one-shot view planning approach, with an $\alpha = 5$ setting, delivers nearly identical quality metrics in PSNR and SSIM when compared to PRV, yet it benefits from reduced movement cost. Moreover, when α is adjusted to 6, our method surpasses PRV in terms of PSNR and SSIM quality while still maintaining a slightly lower movement cost. This confirms that our object-specific view configuration is superior to fixed view configurations in PRV for handling varying geometries of objects.

In conclusion, our RGB-based one-shot view planning method demonstrates several advantages over the baselines. By integrating powerful geometric priors from 3D diffusion models, our method effectively leverages available object information, resulting in more informative and better distributed views. Moreover, our approach showcases superior adaptability through its object-specific view configuration mechanism. Unlike the fixed view configuration in PRV, our method dynamically adjusts the view configurations for different objects based on their varying geometries. However, we observe a longer inference time of our method compared

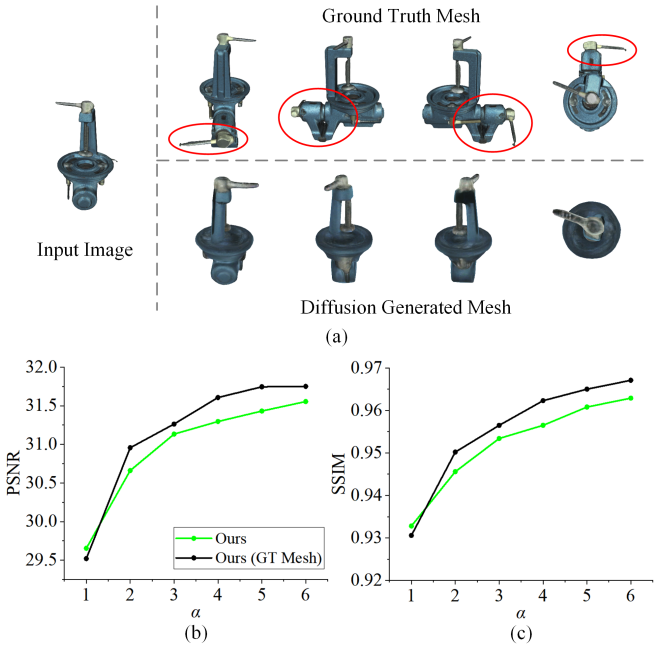

Fig. 8: Analysis of a failure case. Top: Input image to the diffusion model and the generated mesh observed from different perspectives. Red circles indicate the missing parts of the generated mesh compared to the ground truth model. Bottom: We compare the reconstruction results using our one-shot view planning based on the ground truth mesh and the generated mesh, showing that wrongly generated geometry leads to reduced performance.

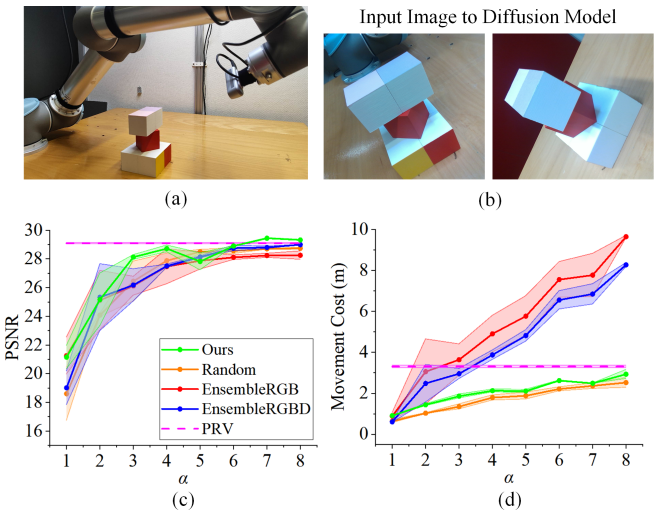

Fig. 9: Real-world experiment showing the test object. We run two test trials with different initial views. PSNR is averaged over 100 novel views. Each value is reported as the averaged mean and with standard deviation (the error bar) on two test trials. By adapting views based on the object geometries, our method achieves a higher performance with lower movement costs.

to the PRV and *Random* methods, primarily due to the constraints imposed by the generation process of the diffusion model (about 60 s) and the online optimization process (about 80 s). We plan to improve this in the future.

E. Analysis of Failure Case

Although our approach successfully performs one-shot view planning from a single RGB image and achieves promising unknown object reconstruction performance, we observe performance inadequacies in a test case. Specifically, the generated mesh from the 3D diffusion model of the Drill object, as depicted in Fig. 8(a), demonstrates geometrical discrepancies compared to the ground truth. These disparities might stem from the limited information available due to occlusion in a single input image and the insufficient representation of this type of object in the training dataset. To further validate the impact of this issue on the reconstruction, we conducted experiments by replacing the generated mesh with the ground truth mesh. Fig. 8(b-c) reveals that our method using the ground truth mesh achieves higher PSNR and SSIM compared to input with the diffusion-generated mesh. The results indicate that the quality of geometric priors, i.e., the mesh generated from diffusion models, is crucial for our one-shot view planning performance.

F. Real-World Experiments

We deploy our approach in a real world tabletop environment using a UR5 robot arm with an Intel Realsense D435 camera mounted on its end-effector (only the RGB optical camera is activated). MoveIt [1] is employed for robotic motion planning. The accompanying video[1] illustrates the online reconstruction process where $\alpha = 7$.

To validate our findings in Sec. IV-D, we compare our method against baselines in the real world. It is worth noting that due to imperfect camera poses and noise in real world experiments, the pose optimization functionality implemented in Instant-NGP is enabled during our NeRF training. The experimental environment and comparisons are shown in Fig. 9. From the results, we confirm that (1) our method generalizes to real world environments; and (2) our method plans object-specific view configurations according to object geometries to achieve higher PSNR with lower movement costs compared to the PRV and NBV methods. Our method achieves peak performance at $\alpha = 7$, which is larger than the value of 6 determined in Sec. IV-B. This might be caused by the noise in the camera pose and images, making it challenging for view planning tasks. We observe a similar slight performance reduction for the NBV methods.

Nevertheless, when deployed in real-world environments, an estimate of the actual object size is necessary to scale the diffusion-generated models, given that the generated mesh lacks scale information.

V. Conclusions

In this paper, we present a novel one-shot view planning method starting with only a single RGB image of the unknown object to be reconstructed. The proposed method exploits priors from 3D diffusion models as a proxy to the inaccessible ground truth 3D model as the basis for one-shot view planning. We develop a customized variant of

[1] https://youtu.be/EKZPHb5-UZk

the set covering optimization problem tailored for NeRF-based reconstruction, which aims to compute an object-specific view configuration that densely covers the generated mesh from 3D diffusion models. We compute a globally shortest path on this view configuration, corresponding to the minimum travel distance. Our experiments validate that utilizing geometric priors from 3D diffusion models yields more informative views compared to Random and next-best-view methods. When compared to the state-of-the-art RGB-based one-shot baseline, our view planning based on varying object geometries demonstrates better performance compared to a fixed view configuration. The real world experiment suggests the applicability of our method.

REFERENCES

[1] S. Chitta, "MoveIt!: An Introduction," *Robot Operating System (ROS) The Complete Reference*, vol. 1, pp. 3–27, 2016.

[2] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A Universe of Annotated 3D Objects," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] N. Dengler, S. Pan, V. Kalagaturu, R. Menon, M. Dawood, and M. Bennewitz, "Viewpoint Push Planning for Mapping of Unknown Confined Spaces," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[4] L. Gurobi Optimization, "Gurobi Optimizer Reference Manual," 2021.

[5] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees," *Autonomous Robots*, vol. 34, pp. 189–206, 2013.

[6] H. Hu, S. Pan, L. Jin, M. Popović, and M. Bennewitz, "Active Implicit Reconstruction Using One-Shot View Planning," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[7] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An Information Gain Formulation for Active Volumetric 3D Reconstruction," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[8] L. Jin, X. Chen, J. Rückin, and M. Popović, "NeU-NBV: Next Best View Planning Using Uncertainty Estimation in Image-Based Neural Rendering," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[9] A. J. Jonathan Ho and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2020.

[10] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, "HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] K. Lin and B. Yi, "Active View Planning for Radiance Fields," in *Robotics Science and Systems (RSS) Workshop on Implicit Representations for Robotic Manipulation*, 2022.

[12] M. Liu, R. Shi, L. Chen, Z. Zhang, C. Xu, X. Wei, H. Chen, C. Zeng, J. Gu, and H. Su, "One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion," *arXiv preprint arXiv:2311.07885*, 2023.

[13] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[14] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-Shot One Image to 3D Object," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[15] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "SyncDreamer: Generating Multiview-consistent Images from a Single-view Image," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2024.

[16] X. Long, C. Lin, P. Wang, T. Komura, and W. Wang, "SparseNeuS: Fast Generalizable Neural Surface Reconstruction from Sparse Views," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2022.

[17] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3D: Single Image to 3D Using Cross-Domain Diffusion," *arXiv preprint arXiv:2310.15008*, 2023.

[18] M. Mendoza, J. I. Vasquez-Gomez, H. Taud, L. E. Sucar, and C. Reta, "Supervised Learning of the Next-Best-View for 3D Object Reconstruction," *Pattern Recognition Letters*, vol. 133, pp. 224–231, 2020.

[19] R. Menon, T. Zaenker, N. Dengler, and M. Bennewitz, "NBV-SC: Next Best View Planning based on Shape Completion for Fruit Mapping and Reconstruction," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020.

[21] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding," *ACM Trans. on Graphics*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[22] S. Oßwald, M. Bennewitz, W. Burgard, and C. Stachniss, "Speeding-Up Robot Exploration by Exploiting Background Information," *IEEE Robotics and Automation Letters (RA-L)*, vol. 1, no. 2, pp. 716–723, 2016.

[23] E. Palazzolo and C. Stachniss, "Effective Exploration for MAVs Based on the Expected Information Gain," *Drones*, vol. 2, no. 1, pp. 59–66, 2018.

[24] S. Pan and H. Wei, "A Global Max-Flow-Based Multi-Resolution Next-Best-View Method for Reconstruction of 3D Unknown Objects," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 714–721, 2022.

[25] S. Pan and H. Wei, "A Global Generalized Maximum Coverage-Based Solution to the Non-Model-Based View Planning problem for object reconstruction," *Journal of Computer Vision and Image Understanding (CVIU)*, vol. 226, p. 103585, 2023.

[26] S. Pan, H. Hu, and H. Wei, "SCVP: Learning One-Shot View Planning via Set Covering for Unknown Object Reconstruction," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 1463–1470, 2022.

[27] S. Pan, H. Hu, H. Wei, N. Dengler, T. Zaenker, and M. Bennewitz, "Integrating One-Shot View Planning with a Single Next-Best View via Long-Tail Multiview Sampling," *arXiv preprint arXiv:2304.00910*, 2023.

[28] S. Pan, L. Jin, H. Hu, M. Popović, and M. Bennewitz, "How Many Views Are Needed to Reconstruct an Unknown Object Using NeRF?" in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[29] X. Pan, Z. Lai, S. Song, and G. Huang, "ActiveNeRF: Learning Where to See with Uncertainty Estimation," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2022.

[30] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D Diffusion," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2023.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[32] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "MVDream: Multi-view Diffusion for 3D Generation," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2024.

[33] N. Sünderhauf, J. Abou-Chakra, and D. Miller, "Density-Aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[34] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[35] Z. Yang, Z. Ren, M. A. Bautista, Z. Zhang, Q. Shan, and Q. Huang, "FvOR: Robust Joint Shape and Pose Optimization for Few-View Object Reconstruction," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] T. Zaenker, J. Rückin, R. Menon, M. Popović, and M. Bennewitz, "Graph-Based View Motion Planning for Fruit Detection," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[37] R. Zeng, W. Zhao, and Y.-J. Liu, "PC-NBV: A Point Cloud Based Deep Network for Efficient Next Best View Planning," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.