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 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, geometric priors about the object are required to conduct one-shot view planning. In this work, we propose a novel one-shot view planning approach that utilizes the 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 [1, 33, 34]. 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 [5, 16, 17, 21, 22, 23, 27, 31, 35]. 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 [4, 24, 26]. 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. A robot's sensor then follows the planned path to collect measurements, which are used for object reconstruction. By decoupling data collection and object reconstruction, these approaches do not rely on iterative map updates for adaptive view planning online. 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 [26] or rely on depth sensors [4, 24].

To address these aforementioned limitations, we propose integrating geometric priors from 3D diffusion models into one-shot view planning. Recently, 3D diffusion models emerge 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 [11, 14, 15]. 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 for 3D reconstruction required in robotics tasks.

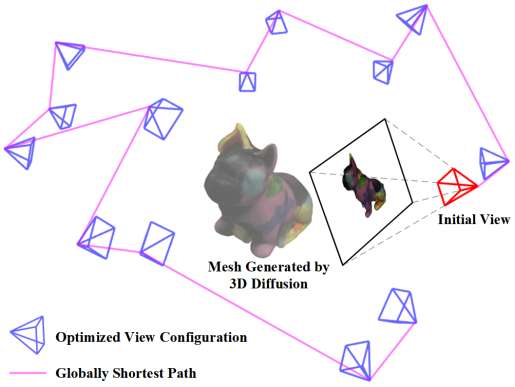

Fig. 1: An example of our RGB-based one-shot view planning by exploiting priors from 3D diffusion models. Our goal is to plan a set of views (blue) at once to collect informative RGB images for object reconstruction. The key component in our approach is a 3D diffusion model generating the corresponding 3D mesh of a single RGB image from the initial camera view (red). By leveraging the mesh as a geometric prior, 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.

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. Given the generated 3D mesh, we convert the one-shot view planning into a customized set covering optimization to calculate the minimum set of views that densely covers the mesh, which we solve using linear programming. 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 conduct extensive experiments on publicly available object datasets and in real world scenarios, demonstrating the applicability of our approach. Our implementation is open-sourced at: `https://github.com/psc0628/DM-OSVP`

*These authors contributed equally to this work. All authors are with the University of Bonn. This abstract is a short version of IROS2024 submission.

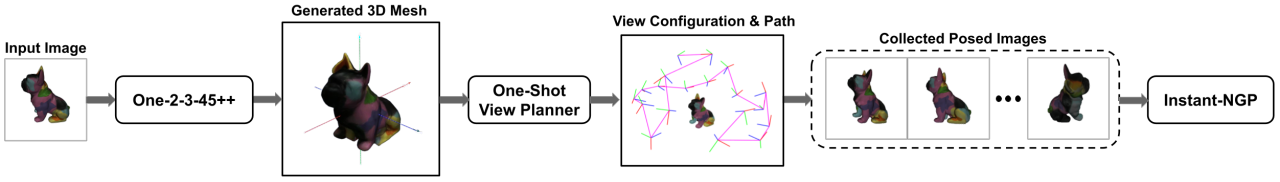

Fig. 2: Overview of our proposed RGB-based one-shot view planning pipeline. Given a single RGB image of the object to be reconstructed, we leverage a state-of-the-art 3D diffusion model, One-2-3-45++ [10], to generate a 3D mesh due to its accurate mesh generation and efficient inference compared to other 3D diffusion models [28, 30, 32]. Based on this prior, we construct the one-shot view planning task as a customized set covering optimization and solve it to obtain a minimum set of views required to densely cover the mesh surfaces. 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 [20]. The RGB camera starts at the initial view and follows the generated globally shortest path to collect RGB images. After data collection, we train a NeRF using Instant-NGP [19] to acquire the final 3D representation of the object.

II. Related Work

View Planning for Object Reconstruction. 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 greedily. In the context of NBV planning for RGB-based reconstruction, Jin et al. [6] integrate uncertainty estimation into image-based neural rendering to guide NBV selection in a mapless way. Lin et al. [9] and Sünderhauf et al. [31] train an ensemble of NeRFs to measure uncertainty for NBV planning.

To improve inefficient path generation and map update of NBV methods, recent works propose the one-shot view planning paradigm. SCVP [24] trains a neural network in a supervised way to directly predict the global view configuration given initial point cloud observations. Hu et al. [4] further reduce the required views by incorporating a point cloud-based implicit surface reconstruction method. In the domain of RGB-based object reconstruction, Pan et al. [26] propose a view prediction network to predict the number of views to reconstruct an object using NeRFs required to reach its peak performance. However, due to the lack of geometric representations during the view planning stage, this work only considers distributing the views following a fixed pattern. Different from previous works that rely on depth sensors [4, 24] or fixed view configurations [26], our novel approach only requires RGB inputs and plans view configurations specifically associated with the objects.

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 [7, 29].

Inspired by the advances of diffusion models, recent works investigate using diffusion models for 3D content generation. They consider fine-tuning pretrained 2D diffusion models for multi-view synthesis from a single image [12, 13]. One-2-3-45 [11] 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++ [10] mitigates the 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 produce high-quality meshes in a short time, i.e., within 60 s.

III. Our Approach

An overview of our approach is shown in Fig. 2. We introduce our one-shot view planner by defining the customized set covering optimization problem below.

A. One-Shot View Planning as Set Covering Optimization

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 [3] 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 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. Indicator function $I(p, v)$ is defined to represent whether a surface point p is observable from view v:

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

Given $\mathcal{P}_{surf}$ and each $\mathcal{P}_v$, the vanilla set covering optimization problem aims to find the minimum set of views required for completely covering the surface points. It requires that each surface point should be covered by at least one view. This definition aligns well with object reconstruction employing depth-sensing modalities [4, 24, 25], 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

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 of 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 $\alpha = 5$. Conversely, the dagger symbol ($\dagger$) indicates non-significant results for $\alpha = 7$ compared to $\alpha = 6$. These are 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.

in our optimization problem, resulting in solutions with more views required. 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. These 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 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:

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

where the objective function $\sum_{v \in V} x_v$ 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.

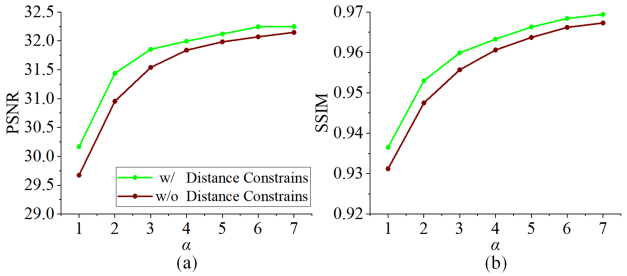

Fig. 3: Ablation study on distance constraints. PSNR and SSIM averaged over 100 novel views. Each value is reported as the averaged mean on 10 test objects. We observed statistically significant results for our method when compared to the version without distance constraints across all α values, as determined 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.

IV. Experimental Results

In simulation, we consider an object-centric hemispherical view space with 144 uniformly distributed view candidates. We set the view space radius to 0.3 m. We test approaches on 10 geometrically complex 3D object models from the HomebrewedDB dataset [8]. 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. We employ the Gurobi optimizer, a linear programming solver [2], to compute the solution for the set covering optimization. To evaluate reconstruction quality, we report peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [18]. We evaluate reconstruction efficiency in terms of inference time for view planning and accumulated movement cost in Euclidean distance.

A. Analysis on Multi-View and Distance Constraints

We first explore the influence of multi-view constraints introduced in Sec. III-A. We test our methods across varying α values from 1 to 7, as detailed in TABLE I. The outcomes justify our modification of the set covering optimization to account for RGB-based object reconstruction using NeRFs.

We next investigate the impact of the distance constraints introduced in Sec. III-A. We adopt binary search in our implementation to find out the object-specific maximum β that still yields a feasible optimization solution. The search

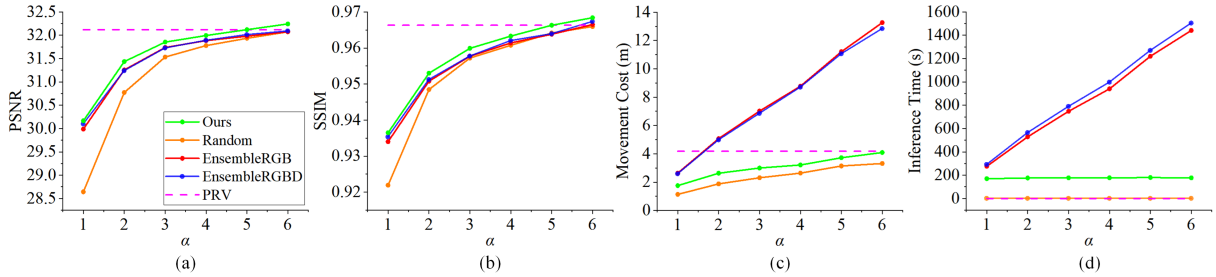

Fig. 4: Comparison to baselines on view planning performance under different α values corresponding to the number of optimized views. PSNR and SSIM are averaged over 100 novel views. Each value reports the mean on 10 test objects. PRV is not associated with α values and is represented by a dashed line. 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 adaptive view configuration is more suitable for object-specific 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 ($\alpha = 6$).

step is set to 0.1 for all experiments. Fig. 3 shows the differences between optimization with and without the proposed constraints over different α values. The results justify our design choice of introducing the distance constraints to find better view configurations.

B. Evaluation of View Planning for Object Reconstruction

Baselines. We compare our novel one-shot view planning with two one-shot baselines (*Random* with globally shortest path and *PRV* [26]) and two NBV baselines (*EnsembleRGB* [9] and *EnsembleRGBD* [31]). As depicted in TABLE I, varying α values lead to 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, excluding PRV, which predicts its own required number of views.

Comparison to Random Selection. As shown in Fig. 4, our RGB-based one-shot view planning approach surpasses the one-shot Random baseline across all α values in terms of PSNR and SSIM. These findings confirm that leveraging powerful geometric priors from 3D diffusion models significantly benefits RGB-based one-shot view planning.

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. 4. We attribute the significant reductions in movement cost and inference time to global path planning and the one-shot paradigm, which avoids iterative map updates and uncertainty computation.

Comparison to PRV. As shown in Fig. 4, our approach with $\alpha = 5$ 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. The results confirm that our adaptive object-specific view configuration is superior to fixed view configurations in PRV for handling varying geometries of objects.

Nevertheless, our method yields longer inference time compared to the PRV and random methods, primarily due to the constraints imposed by the generation process of the diffusion model (approximately 60 s) and the online optimization process (approximately 80 s).

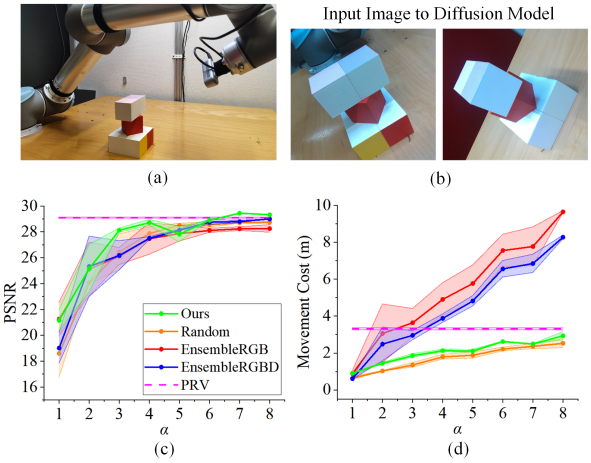

Fig. 5: Real-world experiment showing the test object. We run two test trials with different initial views. Due to imperfect camera poses and noise, the pose optimization functionality implemented in Instant-NGP is enabled during our NeRF training. PSNR and SSIM are averaged over 100 novel views. Each 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 PSNR with lower movement costs.

C. 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). The experimental environment and comparisons are shown in Fig. 5. The demo video can be accessed at: `https://youtu.be/EKZPHb5-UZk`. Our method achieves peak performance at $\alpha = 7$, which is larger than the value of 6 determined in Sec. IV-A. This might be caused by the noise in the camera pose and images, making it challenging for view planning. Nevertheless, when deployed in the real world, an estimate of the actual object size is necessary to scale the diffusion-generated models.

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. Our experiments validate that utilizing geometric priors from 3D diffusion models enables effective RGB-based one-shot view planning.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.