Active View Planning for Radiance Fields

Kevin Lin
UC Berkeley
kevinlin0@berkeley.edu

Brent Yi
UC Berkeley
brentyi@berkeley.edu

***Abstract*—We motivate, discuss, and present a study on the problem of *view planning* for radiance fields. While implicit representations like radiance fields have demonstrated significant promise as a 3D representation for downstream tasks in manipulation, mapping, and navigation, success relies heavily on the coverage of captured views, which are typically manually specified. Our contributions focus on intelligently selecting these views: building on a rich history of classical work in active vision, we (1) discuss the design of `active-3d-gym`, our high-level interface for benchmarking view planning for radiance fields, and (2) propose and experimentally evaluate a simple solution for the view planning problem based on *radiance field ensembles*.**

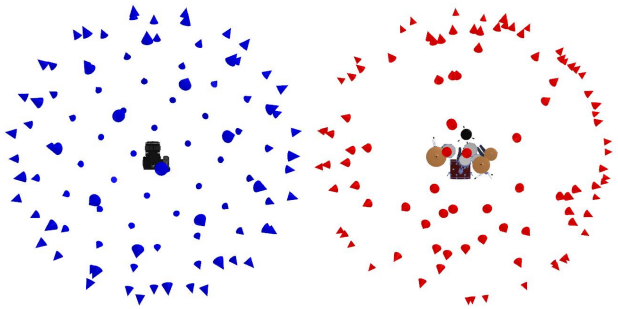

Fig. 1: Visualization of candidate views that can be captured in our `active-3d-gym` environment, with an environment containing an object model imported from ShapeNet [1] and the `nerf-synthetic` [2] dataset. The goal of active view planning is to build a reconstruction of the camera model using a minimum number of these views.

I. Introduction

Recovering 3D representations of objects and scenes from 2D observations is a fundamental task for a range of interests in vision and robotics. Significant progress in this area has recently been made in the form of radiance fields [2], which have enabled a flurry of spectacular results in 3D reconstruction and perception by combining the algorithmic prior of ray marching-based differentiable rendering with an expressive and smoothly optimizable underlying representation.

When applied to robotic systems that need to perceive, interact with, or manipulate unknown environments, however, these techniques are limited by their dependence on the coverage of selected views. While views for the offline datasets typically used for the evaluation and validation of these methods can be manually or densely chosen, it is not always clear what views need to be captured in the real world.

As algorithmic improvements compound with computational ones to enable real-time applications of implicit 3D representations like radiance fields in robotics and manipulation, it will become increasingly important for robots to understand not only how to perceive captured views of unknown environments, but also what views should be captured to begin with. In this direction, we build on a broad body of work in active vision [3] to discuss both a benchmark suite, `active-3d-gym`, and a view planning method, *radiance field ensembles*, to address a simple question: how do we intelligently choose views that result in accurate radiance fields?

II. Related Work

A. Radiance Fields

The success of volume rendering via radiance fields [2] has led to a flurry of recent work in 3D reconstruction and novel view synthesis. Extensions have been abundant, and include addressing of aliasing [4], 360° scenes [5], deformation [6], relighting [7], in-the-wild photos [8], depth perception [9], semantic labeling [10, 11, 12], and uncertainty estimation [13].

While standard neural radiance fields [2] are too slow for real-time use cases, considerable progress has been made in runtime-focused modifications. Several methods have been proposed for distilling pretrained NeRF models into structures amenable to real-time rendering, for example, via caching [14], sparse voxel grids [15], sparse voxel octrees [16], ensembles of smaller networks [17], or planar structures [18]. Other works have improved training time via depth-based supervision [19] or pre-initialization [20], as well as both training and rendering time by combining various spatial data structures with trilinear interpolation [21, 22, 23, 24, 25]. For experiments in this work, we rely on an open-source implementation [26] of Müller et al. [25]'s `instant-ngp`, which enables close-to-real-time training of a range of neural implicit models.

B. Radiance Fields in Robotics

Several recent papers have demonstrated promising results using implicit models like radiance fields in robotics. These include vision-based navigation and state estimation [27], mapping [28, 29], and several works in manipulation: Ichnowski et al. [30] use NeRF's ability to model non-Lambertian surfaces to grasp transparent objects, Driess et al. [31] use object-level NeRF representations for latent dynamics predictions and planning, Li et al. [32] combine NeRF with time-contrastive learning to produce 3D scene representations for

Fig. 2: **Examples of simple environments in `active-3d-gym`.** We provide a shared interface for models imported from a wide range of sources, including ShapeNet [1], `nerf-synthetic` [2], and Replica [45].

manipulation involving both rigid bodies and fluids, and Yen-Chen et al. [33] use NeRF to generate labels for learning dense object descriptors. Our work is inspired by two observations: **(1)** implicit representations will only become more useful for robotics as runtime characteristics are improved by advances in software and hardware, and **(2)** existing works in robotics that leverage these representations typically use hand-specified sets of views, which cannot be relied on in unknown or unstructured environments.

C. Active View Planning

Prior methods for view planning can be loosely classified as either synthesis methods or search methods [3]. Synthesis methods directly calculate the pose a of next best view under certain system and task constraints, and can be either based on hand-designed models [34] or learned from data [35]. Search-based methods, on the other hand, focus on methods for evaluating candidate views. Typically formulated with an information gain metric, these search-based methods have been explored for tactile perception [36], voxels [37], surfels [38], and point clouds [39], with many works focused specifically in the context of manipulation [40, 41]. In this work, we focus on view planning through a search-based lens.

D. Uncertainty estimation

Uncertainty estimation is intrinsic to the view planning problem for implicit models — in the greedy case, view planning reduces to repeatedly selecting actions that minimize the resulting overall epistemic uncertainty. Several approaches proposed for uncertainty estimation in general neural networks can be applied here, notably based on test-time dropout [42, 43] or deep ensembles [44].

We focus on view selection using a photometric uncertainty estimate from an ensemble of radiance field models, which we also compare against the variational S-NeRF [13] approach that relies on uncertainties as a network output.

III. Benchmarking

To evaluate search-based view selection algorithms for radiance fields, we implemented a benchmark suite inspired by OpenAI Gym [46], which we call `active-3d-gym`[1]. The suite includes high-level interfaces for:

- **Environments.** Several offline datasets are supported out-of-the-box. This includes precaptured views adapted from the ShapeNet [1], `nerf-synthetic` [2], and Replica [45] datasets.
- **Actions.** At each step, environments expect an action $a_t \in \mathrm{SE}(3)$ corresponding to one of a fixed collection of training set views that can be taken.
- **Observations.** From the action (selected view), the environment return a corresponding observation tuple $o_t = (I_t, D_t)$, which contains rendered RGB and, if available, ground-truth distance[2] maps as observation. Evaluations can be run both with and without distance supervision (our current experiments focus on the latter).
- **Evaluation metrics.** We evaluate view selection quality by reporting photometric PSNRs (peak signal-to-noise ratios) for visual accuracy and distance errors for geometric accuracy on a held-out set of validation views. To compare view selection strategies, these metrics can be plotted against captured view counts.

Consistent with most prior work in view planning [3], note that the core limitation of the described benchmark is the focus on a "teleportation" context, where we consider the number of views captured but not how the camera is moved to these views. Following prior applications of general view planning to robot manipulators [40, 41], extensions of `active-3d-gym` could include consideration of factors like kinematic feasibility, motion costs, and collisions, in concrete settings with both holonomic and non-holonomic robot embodiments, and with consideration for potential capture of the intermediate views between candidate poses.

IV. Ensemble-based View Planning

Building on the insight that we can approximate the information gain that results from capturing a particular view by computing the epistemic uncertainty associated with it, we study *radiance field ensembles* (RFE), for capturing the information gain associated with candidate views.

The RFE approach follows the search-based planning paradigm, and uses an ensemble of K radiance field models that are trained in parallel. At each step, we allocate each network a fixed budget of rays and apply that budget to minimize standard photometric (and in the future, optionally distance-based) losses:

$$\min_{\theta_k} \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} \left[||C(r) - \widehat{C}_{\theta_k}(r)||_2^2 \right] \tag{1}$$

where θ_k is the parameter vector for the k-th model in the ensemble, r is a ray, $\mathbf{R}$ is the set of all observed rays, $C(r) \in [0,1]^3$ is a ground-truth color, and $\widehat{C}_{\theta_k}(r) \in [0,1]^3$ is a rendered color. Because models in the ensemble are trained on the same set of views but initialized with a different seed, representations of unobserved or hallucinated portions of the scene should differ between the models, while observed

[1]Still under development: https://github.com/kevin-thankyou-lin/active-3d-gym

[2]For simplicity, we standardize on using *distance* rather than *depth* where possible. Given a ray, rendered distance depends only on the location of the ray's origin (that is, the camera's optical center), while depth is a local Z coordinate that also depends on the camera's orientation.

Objects	Lego		Drums		Ficus		Mic		Ship		Chair	
Views	10	20	10	20	10	20	10	20	10	20	10	20
random	17.8 ±1.0	23.5 ±0.7	15.7 ±0.2	18.3 ±0.2	19.1 ±0.1	22.4 ±0.1	21.0 ±0.6	24.9 ±0.5	16.5 ±0.4	19.8 ±0.5	19.7 ±0.9	24.3 ±0.4
even	17.8 ±0.6	23.6 ±0.6	15.8 ±0.1	18.3 ±0.2	19.0 ±0.2	22.6 ±0.1	21.0 ±0.7	25.8 ±0.6	16.6 ±0.2	20.3 ±0.6	19.8 ±0.6	24.7 ±0.5
S-NeRF*	17.2 ±0.7	22.5 ±1.8	14.9 ±0.1	17.1 ±0.0	**19.2** ±0.5	21.8 ±1.0	**23.3** ±1.6	26.1 ±0.7	15.9 ±1.6	19.2 ±1.1	18.1 ±0.0	22.2 ±0.0
RFE	**18.9** ±0.2	**25.4** ±0.3	**16.2** ±0.3	**19.1** ±0.2	18.9 ±0.1	**23.1** ±0.1	21.8 ±0.5	**27.1** ±0.4	**18.1** ±0.2	**23.2** ±0.2	**20.7** ±0.3	**25.7** ±0.1

TABLE I: **Photometric validation PSNRs**, on 10 and 20 captured views in the `nerf-synthetic` [2] environments.

Objects	Bag		Camera		Car		Guitar		Jet		Motorbike	
Views	10	20	10	20	10	20	10	20	10	20	10	20
random	27.4 ±0.4	33.7 ±0.1	24.1 ±0.0	29.2 ±0.2	23.7 ±0.8	**28.1** ±1.4	31.7 ±0.3	35.1 ±0.4	26.4 ±0.3	30.4 ±0.2	26.2 ±0.5	31.0 ±0.6
even	27.5 ±0.2	33.9 ±0.5	**24.5** ±0.9	30.2 ±0.3	**24.0** ±0.6	27.6 ±0.8	**31.8** ±0.2	35.9 ±0.6	26.8 ±0.9	30.6 ±0.5	26.4 ±0.6	31.2 ±0.2
S-NeRF*	17.8 ±6.7	23.7 ±5.9	19.3 ±2.5	24.0 ±0.5	18.3 ±4.2	21.8 ±5.1	28.1 ±0.9	31.3 ±2.6	22.7 ±6.4	27.6 ±1.3	20.3 ±4.2	28.0 ±1.9
RFE	**27.6** ±2.0	**34.9** ±0.7	24.1 ±1.2	**30.8** ±0.5	23.3 ±0.2	27.9 ±1.3	30.8 ±2.3	**36.2** ±0.6	**27.3** ±1.5	**31.1** ±0.2	**26.5** ±0.5	**31.8** ±0.4

TABLE II: **Photometric Validation PSNRs**, on 10 and 20 captured views in environments adapted from ShapeNet [1].

Objects	Bag		Camera		Car		Guitar		Jet		Motorbike	
Views	10	20	10	20	10	20	10	20	10	20	10	20
random	**2.3** ±0.3	**1.6** ±0.1	3.0 ±0.2	**1.9** ±0.1	6.7 ±1.6	**4.3** ±0.4	**2.6** ±0.0	2.0 ±0.4	3.4 ±0.2	**1.9** ±0.0	**3.4** ±0.3	**2.1** ±0.1
even	2.3 ±0.2	1.6 ±0.0	**2.9** ±0.2	1.9 ±0.1	**6.6** ±1.8	4.3 ±0.2	2.6 ±0.2	**1.9** ±0.3	3.2 ±0.0	1.9 ±0.1	3.5 ±0.4	2.1 ±0.1
S-NeRF*	9.4 ±10.3	3.7 ±2.1	3.0 ±0.1	2.6 ±0.0	8.7 ±5.9	5.6 ±2.3	3.7 ±0.9	2.8 ±0.8	10.6 ±13.1	2.6 ±1.2	7.5 ±5.3	2.8 ±1.1
RFE	2.7 ±0.2	1.6 ±0.1	3.0 ±0.2	1.9 ±0.1	7.6 ±0.5	4.4 ±0.1	3.1 ±0.6	2.0 ±0.3	**3.1** ±0.5	2.0 ±0.1	4.3 ±0.5	2.1 ±0.1

TABLE III: **Distance validation errors**, in percent, on 10 and 20 captured views in environments adapted from ShapeNet [1].

portions that are corresponded across observed views will agree.

Based on this intuition, we break RFE down into three steps, which can be repeated until a view budget is expended:

1) **View selection.** For each view a_t of a set of candidate views, we select a sparse set of rays $\mathbf{R}(a_t)$ and render them from each model in the ensemble. Then, we use photometric variances across the ensemble to compute an information gain $IG(a_t)$ for each candidate view:

$$IG(a_t) \propto \sum_{r \in \mathbf{R}(a_t)} \sum_{i=1}^{3} \log(\mathrm{Var}_K(\hat{C}_\theta(r)^{(i)}))$$

where Var_K denotes the variance across the rendered color values for a particular ray across K models in the ensemble. In practice, we find that this formulation can also be successfully implemented as a simple sum over sample variances.
2) **Sensor scanning.** We choose the view with the highest estimated information gain as the next action.
3) **Representation update.** We capture the selected view, and train each model in the ensemble of radiance fields using all views captured thus far.

Although ensemble training is often viewed as expensive [13], in the view selection setting we note that they are actually highly practical from an efficiency standpoint. Ensembles are trivial to parallelize, require neither full convergence of individual radiance fields nor dense rendering of evaluated views, and can take advantage of recent improvements in radiance field training speed [19, 25].

V. Experiments

We use the `active-3d-gym` benchmark suite to compare RFE against three baselines: (1) **random**: simple uniform sampling with replacement, (2) **even**: uniform sampling without replacement, and (3) **S-NeRF***: an uncertainty estimation implementation based on the stochastic NeRF [13] framework.

We report interquartile means and standard errors for different metrics, methods, and environments in Tables I, II and III, for 10 and 20 captured views. Results use ensemble sizes of $K = 2$ and a shared `instant-ngp` [25] architecture for reconstruction and evaluation. The S-NeRF experiments uses a TensoRF [24]-based re-implementation of the S-NeRF formulation (denoted S-NeRF*), but only for uncertainty estimation and view selection; for fairness of evaluation, the selected views are then used to train the same `instant-ngp` architecture that are used for the other experiments.

VI. Discussion

Despite its simplicity, we find that RFE consistently selects views that improve photometric PSNRs more effectively than baselines. Perhaps because it uses photometric disagreement as a proxy for the information gain of a view, however, RFE appears to select views at the cost of distance error.

Consistent with prior work [13], we observed that S-NeRF* is effective in producing high quality uncertainty estimates when provided a large number of input views; our implementation, however, fails to generalize to situations where only a small number of views are available, as is the case in the view planning setting. S-NeRF* underperforms compared to random sampling strategies.

These insights, obtained via `active-3d-gym`, highlight the utility of RFE as a method to tackle the view planning problem for radiance fields. We hope to extend RFE to plan views for radiance fields in physical robot embodiments.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[3] Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3):225–245, Sep 2020. ISSN 2096-0662. doi: 10.1007/s41095-020-0179-3. URL https://doi.org/10.1007/s41095-020-0179-3.

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021.

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *arXiv*, 2021.

[6] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[7] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021.

[8] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021.

[9] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021.

[10] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.

[11] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.

[12] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022.

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

[14] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021.

[15] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.

[16] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021.

[17] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.

[18] Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. Neurmips: Neural mixture of planar experts for view synthesis. *arXiv preprint arXiv:2204.13696*, 2022.

[19] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[20] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *arXiv preprint arXiv:2201.08845*, 2022.

[21] Alex Yu and Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks, 2021.

[22] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33: 15651–15663, 2020.

[23] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, 2021.

[24] Anpei Chen, Zexiang Xu, Andreas Geiger, , Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields, 2022.

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv:2201.05989*, January 2022.

[26] Jiaxiang Tang. Torch-ngp: a pytorch implementation of instant-ngp, 2022. https://github.com/ashawkey/torch-ngp.

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

[28] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew Davison. iMAP: Implicit mapping and positioning in real-time. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[29] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. *arXiv preprint arXiv:2112.12130*, 2021.

[30] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-neRF: Using a neural radiance field to grasp transparent objects. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=zOjU2vZzhCk.

[31] Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields. doi: 10.48550/ARXIV.2202.11855. URL https://arxiv.org/abs/2202.11855.

[32] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. *Conference on Robot Learning*, 2021.

[33] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. NeRF-Supervision: Learning dense object descriptors from neural radiance fields. In *IEEE Conference on Robotics and Automation (ICRA)*, 2022.

[34] Joseph E Banta, LR Wong, Christophe Dumont, and Mongi A Abidi. A next-best-view system for autonomous 3-d object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30 (5):589–598, 2000.

[35] Miguel Mendoza, J Irving Vasquez-Gomez, Hind Taud, L Enrique Sucar, and Carolina Reta. Supervised learning of the next-best-view for 3d object reconstruction. *Pattern Recognition Letters*, 133:224–231, 2020.

[36] Edward Smith, David Meger, Luis Pineda, Roberto Calandra, Jitendra Malik, Adriana Romero Soriano, and Michal Drozdzal. Active 3d shape reconstruction from vision and touch. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16064–16078. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/8635b5fd6bc675033fb72e8a3ccc10a0-Paper.pdf.

[37] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484, 2016. doi: 10.1109/ICRA.2016.7487527.

[38] Riccardo Monica and Jacopo Aleotti. Surfel-based next best view planning. *IEEE Robotics and Automation Letters*, 3(4):3324–3331, Oct 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2852778.

[39] Rui Zeng, Wang Zhao, and Yong-Jin Liu. Pc-nbv: A point cloud based deep network for efficient next best view planning. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7050–7057, 2020.

[40] Michael Krainin, Brian Curless, and Dieter Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation*, pages 5031–5037. IEEE, 2011.

[41] Jinda Cui, John T Wen, and Jeff Trinkle. A multi-sensor next-best-view framework for geometric model-based robotics applications. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8769–8775. IEEE, 2019.

[42] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[43] Sergio Hernández, Diego Vergara, Matias Valdenegro-Toro, and Felipe Jorquera. Improving predictive uncertainty estimation using dropout–hamiltonian monte carlo. *Soft Computing*, 24(6):4307–4322, 2020.

[44] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

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

[46] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.