Deep Sharpness-Aware Next Best View Selection for Grided View Synthesis

Anonymous CVPR submission

Paper ID 16110

Abstract

Next Best View (NBV) aims at identifying the next most informative sensor position (view) for 3D reconstruction or view synthesis in a 3D scene. In this paper, we focus on NBV for view synthesis with the emerging representation of implicit voxel grids, which is becoming the game changer with its advantages to recover fine details of the 3D scene while significantly saving training time and training GPU memory footprint. However, the nature that the latent codes in each voxel cell on the voxel grid are separately and imbalancedly trained induces problem with the existing prediction-based NBV methods where no reasonable information gain prediction can be made across different cells. To overcome this problem, we present SharpView, an NBV algorithm that first introduces the sharpness of loss space into information gain estimation based on an intuition that the training model has a soft output space around the finely synthesized views. To estimate sharpness of loss space of a candidate view, we design a pseudo labelling mechanism that incorporates the output of previous trained views and estimate the gradient embedding norm in the last model layer. We conduct experiments on various view synthesis benchmarks which confirmed that SharpView outperforms the baselines in finding NBV for view synthesis with the representation of implicit voxel grids. The source code for result reproduction is available at `https://github.com/cvpr_16110/SharpView`.

1. Introduction

Next Best View (NBV) [4, 7, 10, 13] aims to iteratively find a shortest and most informative sequence of sensor positions (views) to acquire RGB images in a previous unknown scene to boost efficiency and accuracy of view synthesis of 3D scenes [13, 17, 24], which is a fundamental task for downstream applications such as augmented/virtual reality and autonomous driving [6]. The existing NBV methods [13, 21, 22] for classic view synthesis typically feature the model representation of a sole multi-layer perception model (MLP) and predicting output uncertainty (e.g., standard deviation) from a view.

However, while the emerging representation of implicit voxel grid for view synthesis (grided view synthesis) [17, 24, 28] is becoming a game changer of this domain with its advantages to recover fine details of the 3D scene while significantly saving training time and training GPU memory footprint, it is also incurring problems for the existing NBV methods [13, 21, 22] for view synthesis. Specifically, such representation divides the 3D scene and the training model and map them into pairs of local geometries and local latent codes in grid cells. And each local latent code is trained to model comparatively a simpler local geometry for fast convergence and only a subset of local latent codes need to be optimized for each view.

In such case, each voxel cell on the voxel grid are separately and imbalancedly trained and there is often not enough information for a local latent code to model the standard deviation of its output: frequently visited local latent codes will output low uncertainty prediction while those less frequently visited outputs randomly. We even recorded worse NBV selection performance of the prediction-based NBV methods than the random strategy.

To tackle this problem, we notice an intuition that under small view position and direction perturbation, the observed RGB value of a spot in the 3D scene varies in a continuous way. It implies that the output space (and loss space) of pixels of the image captured from a finely synthesized view is flat. Thus the sharpness of the loss space can serve as a hint to find the uncertain views with high information gain, without the need for extra training.

Based on the above observation, we propose SharpView, an NBV algorithm that first introduces the sharpness of loss space into information gain estimation in NBV selection for grided view synthesis. For estimation of the sharpness of loss space of a view, we generate reasonable pseudo labels (RGB values) by referring the output of the training model from the previously trained views at the same spot as a perturbation for loss computation. We then back propagate the computed loss against the pseudo labels and compute the sharpness of loss space of a view as the norm of last layer of parameters of the model (the smaller MLP).

In this way, we are able to find the most uncertain views with sharpest loss space, without being influenced by imbalanced training progress of each voxel cell. We summarize our contributions as follows:

- SharpView is the first NBV algorithm that takes sharpness of loss space into consideration for information gain estimation in NBV selection.
- We design a pseudo labelling mechanism for new views for sharpness computation.
- Sharpness performs best among the baselines in various benchmarks.

We discuss the related work in Section 2, and introduce our notation, settings and algorithm of SharpView in Section 3. We present our experiments in Section 4 and then conclude in Section 5.

2. Background

2.1. Neural Implicit Representations

Neural implicit representation [13, 17, 19, 23, 28] is a emerging mapping representation demonstrating promising results for object geometry reconstruction, scene completion, novel view synthesis and also generative modelling. They typically feature a Neural Radiance Field (NeRF) [19] structure that learns a density and radiance field supervised by 2D views (camera position & orientation and the images captured) with an MLP model. iMAP [23] uses a single MLP neural model as the underlying 3D scene representation and with a comparatively simple implicit representation and efficient rendering pipeline, iMAP achieves near real-time performance in training.

However, recent researches [5, 9, 11, 17, 28] report a single MLP representation is not scalable due to limited capacity and tends to ignore complex details. They propose to decompose the whole 3D scene to grided local scenes and train local implicit representations to map the local geometry in each local scene, which improves the level of detail in reconstruction because each local implicit representation only needs to map a local region rather than the geometry of a whole scene. Organizing the local implicit representations as a grid, we can easily find the 3D correspondence between the local implicit representations and the 3D scene, which is the basis of our proposed method.

2.2. Grided Implicit Rendering

Along the decomposition of the implicit representation to grided local implicit representations, some [17, 28] also decompose the training pipeline to effectively leverage prior knowledge of local geometries embedded in MLP. Specifically, they [17, 28] separate the MLP model to an AutoEncoder-like network that consist of an encoder, grided latent codes and a decoder, with the encoder and the decoder pretrained over various scenes to extract generalizable knowledge of 3D reconstruction. Because they only need to optimize the local latent codes, they manage to reconstruct complex geometry of 3D scenes in a view with several training iterations, retaining real-time performance as imap [23]. Among these work, we select BNV-Fusion [17] as our major research target, which takes depth images and camera poses as input and achieves high quality shape reconstruction of complex 3D scene.

2.3. Next Best View

Traditional NBV [4, 7, 10] typically aims to find a shortest sequence of views from a set of candidate views that optimize the coverage of a previously unknown area. Given the existing partial explicit map (e.g., point cloud), they either heuristically find frontiers of the map, or predict views with AI models that optimize coverage. With the emerging implicit 3D reconstruction being able to reconstruct finer details of the complex 3D scene, optimizing accuracy is also an emerging requirement for NBV [13, 20–22], where the most valued information gain is the improvement of the quality of the reconstructed 3D model.

2.4. Uncertainty Estimation

The information gain from training data for a training model can be directly modeled as the uncertainty reduction of model parameters, and such uncertainty estimation is a long-standing problem [1, 3, 8, 16, 18, 25] for machine learning. A classic framework for uncertainty estimation is the Bayesian Learning framework that estimates the posterior distribution of the model given the existing training data. However, such approaches typically require multiple model evaluations which are computationally expensive, and require significant modifications over network architectures and training procedures [22, 25].

Recent work focusing on the NeRF structure approximate the uncertainty of model parameters with the posterior distribution of the output density and radiance (uncertainty of the model output) [13, 20–22]. They typically follow the pattern of generalization of standard NeRF [22] that learns a probability distribution over all the possible radiance fields modeling the scene, where an extra model head is trained to estimate the variance of the radiance fields under the supervision of existing views.

2.5. Connection to Other Sharpness Methods

In the domain of active learning, there exists other methods [2, 12, 14, 15] that also inspect the flatness of loss space and use gradients as an indicator of its sharpness. To the best of our knowledge, they [2, 12, 14, 15] are focused on classification problems where the model output is the activated by a softmax function and the pseudo labelling policy is simply selecting the value with highest probability in the output, which cannot work with implicit rendering pipeline where rgb values are directly output . We are the first work to connect loss space sharpness with implicit rendering pipeline by introducing the pseudo labelling policy based on the intuition of continuity of rgb values around a small region of viewpoints.

2.6. Grided View Synthesis

Here we discuss the notation of grided view synthesis in details. Assume that we equally divide the 3D scene of interest into M cells under certain resolution and the implicit voxel grid for view synthesis can be represented as pairs of 3D coordinates of local geometries and local latent codes modelling the local geometry: $G = \{(x_i, \theta_i)\}_{i=1}^{M}$. Given a step on a ray $r = (x, d)$ where x is a coordinate in the 3D scene and the d is the viewing direction, we query the pairs for the closest distance $\theta_j = Q(G, r)$ and then use MLP models to decode the density σ and view-direction-dependent color c from the corresponding local latent codes: $\sigma, c = MLP(\theta_j, d)$.

To render the color $C(r)$ of a pixel on an image from a view, we cast a ray with K steps, $\boldsymbol{r} = \{r_i\}_{i=1}^{K}$ from the center of the camera through the pixel and query the implicit voxel grid for density and rgb values at each step $\{\sigma_i, c_i\}_{i=1}^{K}$. Finally, the queried results are accumulated to compute $C(r)$.

$$C(\boldsymbol{r}) = \sum_{i=1}^{K} T_i \alpha_i c_i \tag{1a}$$

$$\alpha_i = 1 - exp(-\sigma_i \delta_i) \tag{1b}$$

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{1c}$$

where α_i is the probability of termination of ray at step i, T_i is the accumulated transmittance and δ_i is the distance between consecutive steps. We gather the rgb value of each pixel on the image from a view and forms the predicted image I from view v and the loss is calculated against the groundtruth image $\hat{I}$ with mean square error $l_{mse}(I, \hat{I})$.

3. Methodology

Given a set of N views $V = \{v_i\}_{i=1}^{N}$ discretized on a sphere, assume the above grided view synthesis pipeline has been trained by a subset $S \subset V$ and their corresponding groundtruth images, and the candidate views for NBV selection is $U = \frac{V}{S}$. We approximate the information gain of a view with the value of loss function of the grided view synthesis pipeline. To solve the NBV problem, we aim to find a view that maximize the loss between rendered image from the view and the corresponding groundtruth image, which typically complies with the sharpest loss space around the view.

3.1. Pseudo Ray Labelling

Algorithm 1: Pseudo Ray Labelling

Input: A ray: $\boldsymbol{r}$; Set of trained views: S; implicit voxel grid: G

Output: Pseudo accumulated rgb value of $\boldsymbol{r}$: $\hat{C}(\boldsymbol{r})$

Compute the dominant step r^d along r;
Compute the closest view $v' \in S$ from r^d;
Cast ray $\boldsymbol{r}'$ from v' to r^d;
Render pseudo rgb value $\hat{C}(\boldsymbol{r})$ of $\boldsymbol{r}'$ from G.

To compute a reasonable pseudo label for loss calculation to estimate sharpness of loss space, we use views in S as references assuming that the implicit voxel grid has already fully learned knowledge from their corresponding images. Assume we are estimating loss space sharpness of a candidate view v and we are casting a ray $\boldsymbol{r}$ with K steps to predict the rgb value of a pixel on the predicted image from view v.

In view of the continuity of rgb value when the viewing direction is slightly perturbed, a reasonable pseudo label for ray $\boldsymbol{r}$ would be the rgb value predicted from the closest trained view, which is also the most certain value that can be predicted from the grid. As shown in Algorithm 1, to compute the closest trained view, we first find a dominant step r^d along $\boldsymbol{r}$:

$$\begin{aligned} r^d &= r_k \\ k &= \underset{0<i\leq K}{\arg\max}\, T_i \alpha_i \end{aligned} \tag{2}$$

α_i and T_i are the probability of termination of ray at step i and the accumulated transmittance from Equation 1a when rendering accumulated rgb value along $\boldsymbol{r}$. We are treating the term $T_i \alpha_i$ as weights of each step and the one with largest weight is the dominant one.

Then we cast rays from all trained views to the coordinate x^d determined by r^d and the ray with smallest intersection angle with r^d is selected as the reference ray, and we render the rgb value $\hat{C}(\boldsymbol{r})$ of this reference ray as the pseudo label for $\boldsymbol{r}$.

3.2. Sharpness Estimation

The procedure of loss space sharpness estimation for a view v in shown in Algorithm 2. Gathering pseudo label of each ray through each pixel on the predicted image I from view v we will get a pseudo image $\hat{I}$. After computing mean square error loss between I and $\hat{I}$ and back propagating the loss, we use the norm of gradients of weights from the last layer of decoder MLP as an estimation of sharpness. After estimating the sharpness among all candidate views in U, the one with highest sharpness value is selected as the next best view v_n. We then acquire its groundtruth image, append to the training set S and remove it from the candidate set U.

Algorithm 2: Loss Space Sharpness Estimation

Input: A candidate view: v; Set of trained views: S; implicit voxel grid: G; MLP decoder: M
Output: Sharpness of loss space around v: s

Render predicted image I of v with G and M;
$\hat{I}$ = copy(I);
for *pixel* $\in \hat{I}$ **do**
    Cast ray $\boldsymbol{r}$ from v to pixel;
    $\hat{C}(\boldsymbol{r})$ = PseudoRayLabelling($\boldsymbol{r}$, S, G);
    Replace pixel value on $\hat{I}$ with $\hat{C}(\boldsymbol{r})$;
Back propagate $g_I = \frac{\partial}{\partial M_{out}} l_{mse}(I, \hat{I})$, where M_{out} is the weight of the final layer;
$s = \|g_I\|_2$.

4. Experiments

4.1. Evaluation Setup

TestBed

We evaluated SharpView on a PC equipped with Nvidia 2080Ti 11GB GPU, Intel i5-12400 CPU and 32GB RAM.

Baselines

We compare SharpView (referred to as Sharp.) against three baselines: random (referred to as Rand.), maximal distance (referred to ass MDist.) and prediction-based NBV (referred to as Pred.). Rand. is a pure randomized strategy. MDist. maximizes the distance between selected view to the view from the training dataset. Pred. is a prediction-based method that trains an extra head on the MLP decoder of the grided view synthesis pipeline following the patterns of loss functions commonly used in these methods [13, 20–22]:

$$L = \frac{1}{R} \sum_{r=0}^{R} \big(\log \sigma_r + \frac{(c_r - \hat{c}_r)^2}{\sigma_r^2} \big) \tag{3}$$

where R is the number of pixels in an input image, σ_r is the standard deviation prediction from the extra model head, c_r is the predicted rgb value, and $\hat{c}$ is the groundtruth rgb measurement.

Workload

We choose DirectGO [24] as the grided view synthesis pipeline for our evaluation, which features a implicit voxel grid representation of the 3D scene and short convergence time within minutes, compared with the convergence time of days using the non-grided counterparts. It also achieves comparable resulting view synthesis accuracy. We follow DirectGO to use PSNR, SSIM [26], and LPIPS [27] as the metrics to compare the view synthesis accuracy among SharpView and the baselines.

Datasets

We evaluate our approach on three different datasets. We configure the datasets mainly following the default setup of DirectGO [24]. Synthetic-NeRF contains eight objects with realistic images synthesized by NeRF. Synthetic-NSVF contains another eight objects synthesized by NSVF. We follow the setup of these two datasets and set the image resolution to 800 $\times$ 800 pixels and set 100 views for training and NBV selection and 200 views for testing for each scene. TanksAndTemples is a real-world dataset captured from large real-world 3D scenes with 1920 $\times$ 1080 image resolution. We set one-eighth of the images testing and the rest for training and NBV selection.

NBV Configuration

We initialize the NBV procedure by training the grided view synthesis pipeline with a initial training dataset sized six views. And the NBV procedure ceases after acquiring ten new views from the training dataset. The six views are selected by randomly selecting the first view and then append the rest views with the same policy with MDist., so that all surfaces of the 3D scene are more likely to be covered at the initial stage. After a view is appended to the training set, we retrain the grided view synthesis pipeline to avoid overfitting to the previous views in the training set and compute new NBV information gain estimation. In the training procedure, we scale the default configuration of number of training iterations and learning rate decay from DirectGO [24] by the ratio between the length of training set and the whole training dataset. At the end of the NBV procedure, the view synthesis accuracy results are calculated and we present below the results averaged over repeating the evaluation with three different seeds.

4.2. Quantitative Comparison

We first qualitatively compare the view synthesis accuracy results under different NBV methods. With the knowledge of loss space sharpness of each candidate views, SharpView managed to select views with more information gain from the candidate views as the NBV, and constantly outperformed the baselines in terms of PSNR, SSIM and LPIPS as shown in Table 1, 2 and 3.

Note that while the results of Rand. and MDist. are comparable since they are both naive methods without any information gain estimation, Pred. constantly performed the worst, which means that Pred. tended to select views with less information gain in the grided view synthesis settings. The possible reason for this phenomenon is two-fold. First, the extra term and factor introduced in their loss function as in Equation 3 in additional to the mean square error may decrease the magnitude of the computed loss value and thus slow down convergence. Second, the local latent codes in different cells are separately and imbalancedly trained, which means the supervision of certain cells can be weak, especially the less frequently visited and thus uncertain cells. This results in that the frequently visited cells output low uncertainty (standard deviation of its rgb output σ) and less frequently visited cells output uncertainty of comparatively random values, severely interfering their NBV selection.

4.3. Qualitative Comparison

Here we present some details of the tested view synthesis results after the NBV procedure of each method for qualitative comparison, which shows that synthesized views with SharpView recovered finer details of the 3D scenes.

4.4. Discussion and Limitation

Due to limited time budget, we did not make it to broadly and extensively evaluate SharpView and we are taking it as the future work. Although we are motivated by the problems induced by grided view synthesis, the resulting algorithm and pipeline does not rely on the architecture or design of grided view synthesis, and we are curious about whether similar advantages can be achieved on other view synthesis architecture or even other domain of implicit rendering such as 3D reconstruction, which is also regarded as our future work.

5. Conclusion

While the cutting-edge grided view synthesis boost the state-of-the-art performance of view synthesis, it incurs difficulty for NBV selection since latent codes in the grid are imbalancedly trained. In this paper, we present SharpView, the first NBV algorithm for grided view synthesis that incorporates loss space sharpness estimation into information gain estimation for NBV selection. By simply leveraging pseudo ray labelling, we estimate the sharpness of loss space of the current training model at candidate views by calculating the gradients of parameters of the last layer of the view synthesis pipeline, without the need to train or infer extra information from the latent codes, so that more accurate information gain estimation can be achieved.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297, 2021. 2

[2] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. 2

[3] Pier Giovanni Bissiri, Chris Holmes, and Stephen Walker. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016. arXiv:1306.6430 [math, stat]. 2

[4] Fredrik Bissmarck, Martin Svensson, and Gustav Tolt. Efficient algorithms for next best view evaluation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5876–5883, 2015. 1, 2

[5] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction, 2020. 2

[6] Shengyong Chen, Youfu Li, and Ngai Ming Kwok. Active vision in robotic systems: A survey of recent developments. 30(11):1343–1377. Publisher: SAGE Publications Ltd STM. 1

[7] C. Connolly. The determination of next best views. In *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, pages 432–435, 1985. 1, 2

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. 2

[9] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape, 2020. 2

[10] Antoine Guedon, Pascal Monasse, and Vincent Lepetit. SCONE: Surface Coverage Optimization in Unknown Environments by Volumetric Integration. *Advances in Neural Information Processing Systems*, 35:20731–20743, 2022. 1, 2

[11] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6000–6009, 2020. 2

[12] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. 2

[13] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. NeU-NBV: Next Best View Planning Using Uncertainty Estimation in Image-Based Neural Rendering, 2023. arXiv:2303.01284 [cs]. 1, 2, 4

[14] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. 2

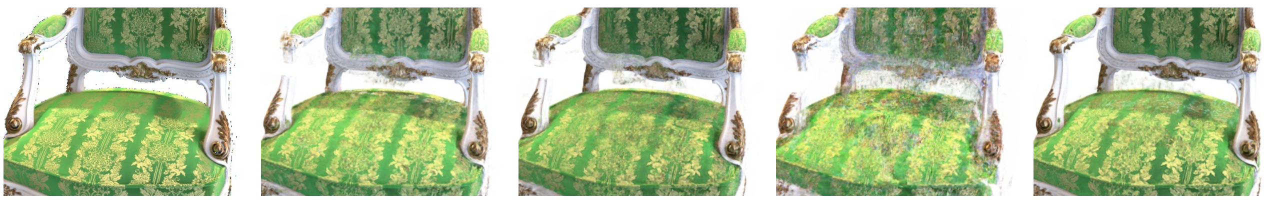
[15] Yoon-Yeong Kim, Youngjae Cho, Joonho Jang, Byeonghu Na, Yeongmin Kim, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. SAAL: Sharpness-aware active learning. In

Table 1. Results on Synthetic-NSVF.

Methods	Palace			Robot			Spaceship		
	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Rand.	27.841	0.884	0.084	25.777	0.949	0.047	26.181	0.944	0.05
MDist.	27.56	0.879	0.082	25.357	0.946	0.049	25.805	0.942	0.05
Pred.	26.111	0.849	0.118	23.454	0.929	0.070	24.049	0.928	0.073
Sharp.	**28.768**	**0.897**	**0.069**	**27.653**	**0.963**	**0.025**	**26.561**	**0.948**	**0.047**

Table 2. Results on Synthetic-NeRF.

Methods	chair			lego			ship		
	PSNR↑	SSIM ↑	LPIPS ↓	PSNR↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Rand.	26.03	0.916	0.082	25.889	0.904	0.062	23.871	0.797	0.192
MDist.	27.034	0.928	0.07	26.392	0.911	0.056	24.609	0.798	0.184
Pred.	23.224	0.871	0.129	22.521	0.842	0.117	22.744	0.747	0.227
Sharp.	**28.573**	**0.936**	**0.049**	**27.133**	**0.919**	**0.047**	**25.116**	**0.805**	**0.171**



(a) groundtruth (b) Rand. (c) MDist. (d) Pred. (e) Sharp.

Figure 1. Qualitative Comparison on Synthesis Nerf chair.

(a) groundtruth (b) Rand. (c) MDist. (d) Pred. (e) Sharp.

Figure 2. Qualitative Comparison on Synthesis Nerf ship.

Table 3. Results on TanksAndTemple (Averaged across different scenes).

Methods	PSNR↑	SSIM ↑	LPIPS ↓
Rand.	21.369	0.871	0.210
MDist.	21.188	0.867	0.217
Pred.	17.437	0.837	0.264
Sharp.	**22.674**	**0.879**	**0.200**

Proceedings of the 40th International Conference on Machine Learning, pages 16424–16440. PMLR. ISSN: 2640-3498. 2

[16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. 2

[17] Kejie Li, Yansong Tang, Victor Adrian Prisacariu, and Philip H.S. Torr. BNV-Fusion: Dense 3D Reconstruction using Bi-level Neural Volume Fusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6156–6165, New Orleans, LA, USA, 2022. IEEE. 1, 2

[18] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning, 2019. arXiv:1902.02476 [cs, stat]. 2

[19] 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, 2020. arXiv:2003.08934 [cs]. 2

[20] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning Where to See with Uncertainty Estima-

tion. In *Computer Vision – ECCV 2022*, pages 230–246. Springer Nature Switzerland, Cham, 2022. Series Title: Lecture Notes in Computer Science. 2, 4

[21] Yunlong Ran, Jing Zeng, Shibo He, Lincheng Li, Yingfeng Chen, Gimhee Lee, Jiming Chen, and Qi Ye. NeurAR: Neural Uncertainty for Autonomous 3D Reconstruction with Implicit Neural Representations, 2023. arXiv:2207.10985 [cs]. 1

[22] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic Neural Radiance Fields: Quantifying Uncertainty in Implicit 3D Representations, 2021. arXiv:2109.02123 [cs]. 1, 2, 4

[23] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. iMAP: Implicit Mapping and Positioning in Real-Time. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6209–6218, Montreal, QC, Canada, 2021. IEEE. 2

[24] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5459. ISSN: 2575-7075. 1, 4

[25] Dustin Tran, Michael W. Dusenberry, Mark van der Wilk, and Danijar Hafner. Bayesian Layers: A Module for Neural Network Uncertainty, 2019. arXiv:1812.03973 [cs, stat]. 2

[26] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 4

[27] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[28] 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. pages 12786–12796, 2022. 1, 2