Online House Touring: Novel View Synthesis on Complex Indoor Scenes

Yilin Wu
yilinwu@stanford.edu

Qinchen Wang
qinchenw@stanford.edu

Jiani Wang
jianiw@stanford.edu

Abstract

Novel view synthesis is a long-standing problem, lying in the intersection of computer vision and computer graphics. Our project focuses on the task of novel view synthesis for indoor scenes. The motivation behind this is that if we can successfully generate novel scenes given only 2d image inputs, we can develop applications such as virtual house showing that only requires a few images of the house through some easily accessible device such as a smartphone. Recent years, deep learning methods such as neural radiance field (or NeRF) has shown great results in novel view synthesis for 3d objects, but the difficulty remains with indoor scene inputs, where different images can have largely varying angles. We chose a total of six indoor scenes, with 4 scenes from the dataset Hypersim, one scene from ScanNet and one custom scene we took with our own iPhone. Each scene consists of 40-60 image inputs with largely varying camera angles. We compared performance between NeRF and NerfingMVS on all these scenes. Experiment result indicates that NeRF has relatively better result on all kinds of indoor scene dataset, including complex and finer geometric dataset. The views generate by NeRF has high quality and perfom better on evaluation metric, which can further generate continous views or video to achieve our goal of online house touring.

1. Introduction

Traditional methods in the field of novel view synthesis can be traced back to 1990s, with early work concentrating on interpolation either between corresponding pixels from the input images, or between rays in space. However, they all suffer from the loss in image quality. A recent breakthrough in the area of novel view synthesis is to apply deep learning techniques to tremendously improve the results, thus gaining renewed popularity. NeRF (Neural Radiance Fields)[18] is such a method that achieves state-of-the-art results for synthesizing novel views of complex scenes. A lot of recent follow-up work[22] on Nerf also extends the usage of Nerf model.

Although there are a variety of scenes suitable for the view synthesis task, we are particularly interested in the indoor scenes because considering the impact of COVID-19 and the need for a house showing before purchase, it is necessary and critical to develop a technique that could synthesize different views of the indoor furniture of a house based on a few images provided by the seller. Therefore, in this project, we are going to explore several existing approaches in novel view synthesis on the dataset(s) of indoor scenes. We would use the traditional view morphing techniques as our baseline and then use Nerf[18] and NerfingMVS[29] as two other methods to generate novel view of a indoor scene. Moreover, considering the different use cases of house showing, we test our methods both on synthetic datasets and real-world datasets. The synthetic datasets include HyperSim[23]and ScanNet[7]. The real-world dataset are the manually captured scenes in the Gates Building. We conduct an extensive study over these models and datasets to approach our problem and analyze the results both qualitatively and quantitatively.

2. Related Work

2.1. Multi-view reconstruction

Multi-veiw reconstruction is an interesting question originated from transitional 3D vision area. Traditional methods use view morphing[25] and stereo vision method to reconstruct images of new views between two different images. Other 3D vision methods [9, 28] focusing on using photo-consistency constrains and optimization methods to reconstruct from images. For example space curving[14] is a widely used, volumetric approach. These methods are based on similarity of image pixels, so they work poorly if the texture of indoor scene are difficult to distinguish and extract feature. Recently, there are a lot of learning and neural network based methods[12, 18, 30] focusing on solving 3D reconstruction task. These methods use similar ideas from conventional method and learning-based optimization, and get good performance on different datasets. Thus, our project mainly want to compare these methods and systematically learn the generalize ability of these methods. We

also aim to find an appropriate method with high-quality novel view synthesis and less training cost that can be applied to people's daily life usage, for example online house touring.

2.2. Neural Implicit Representations

An essential part of a neural network model that can predict novel views is to be able to understand geometry in 3D using some implicit representation - this is referred to as neural implicit representation in the NeRF [18] paper. One line of work focuses on representing 3D geometry as level sets [10, 13, 16, 21]. However, these models require access to ground truth 3D geometry, which is difficult to obtain in commercial settings. Another line of work relaxes this requirement of ground truth 3D geometry, and only requires 2D image inputs [20, 27]. However, these works are limited in their capacity to model complex geometry. NeRF proposes a representation for an input of 5D radiance fields, made up of 3 spacial dimensions and 2 dimensions for camera viewing angle. Building on top of NeRF, NeRF in the wild [15] uses latent appearance rendering such that representation becomes image dependent. This mitigates the limitation of NeRF on scenes with varying lighting and provides insight for possible extensions of NeRF for various scenes.

2.3. View Synthesis

View synthesis aims to create new views of a specific subject starting from a number of pictures taken from given points of views. From images, one can estimate global scene geometry using structure from motion techniques[1, 8, 11]. The estimated geometry can be used to render input images into novel camera viewpoint. One popular class of such approaches are mesh-based representations of scenes including diffuse[5] and view-dependent[8, 2] appearance. However, the generated results are inclined to have artifacts because of the sparsity of the point cloud and it is difficult to get a single consistent global model of entire scene geometry. Therefore, some methods[3, 4] switch to use per view local geometry. Some of them use depth estimation to assist the generation of novel views.

Apart from reconstructing the geometry, view synthesis could also viewed as a problem of view interpolation. One such technique applied to view interpolation is view morphing[26]. We use this as one of our baselines but generally this method fail to render the full scene from a novel viewpoint. Volumetric representation[19] is another traditional computer vision technique that naturally deals with view synthesis because it is able to represent complex shapes and materials realistically and well-suited for gradient-based optimization and tend to produce less visually distracting artifacts then mesh-based methods. Recently, with the success of deep learning, one volumetric approach Nerf[18] has demonstrated impressive results and remarkable improvement in view synthesis. One extension of Nerf[29] tries to use depth information to guide the optimization of the model and to improve the synthesis quality. In the following sections, we are going to study our view synthesis problem based on both Nerf and this Nerf-extended method.

3. Approaches

3.1. Traditional Approaches

View morphing [26] is a classical method to reconstruct 3D scene according to 2D images. In our project, we use view morphing as a baseline method and compare other neural network based methods with it.

View morphing contains following step:

1. Using 8 points algorithm to estimate the fundamental matrix. In this step, we need to get at least 8 corresponding point pairs in two input 2D images.

2. Pre-warp the images, which uses image rectification method. In this step, we find homographs H_1 and H_2 and transform two images.

3. Morphing two images and get the interpolation of two pre-warped images $\hat{I}_1$ and $\hat{I}_2$. For some fraction s, the interpolated image is:

$$\hat{I}_s(x, y) = s\hat{I}_1(x, y) + (1 - s)\hat{I}_2(x, y) \qquad (1)$$

4. Postwarp, this step will transfer the intermediate morphed image back to a specific image plane. We apply homography to intermediate image and get the final views.

The result of traditional view morphing method is shown as figure 1. The view morphing method have several limitations. Firstly, we need to manually choose corresponding points in our indoor scene dataset, which is not accurate and bring a lot of error for further steps. Secondly, it can only generate views from two images, even if other images in dataset might also contains information of this 3D area. Finally, the visualization result shows that the boundary of 3D objects is blur, and the visualization result is not satisfying. So we then chooses and studies neural network based methods to get better result.

3.2. Nerf

NeRF [18] is a popular state of the art rendering technique in generating novel views. It achieves impressive performance in rendering novel views with high resolution. The training of a NeRF model requires a 5D input consisting of 3D (x, y, z) position vector of the object, and a 2D vector (θ, ϕ) representing the camera viewing angle. The

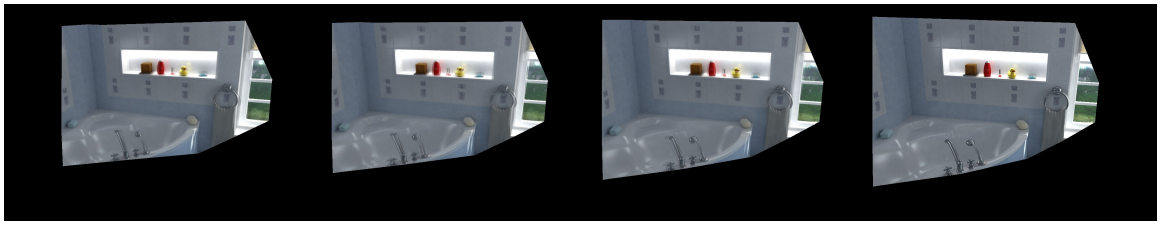

Figure 1. Traditional Method - View morphing result

2D vector represents the (θ, ϕ) values in a spherical coordinate whose origin is at the (x, y, z) of interest. For each viewing angle (or camera ray), 5 samples were taken, and the model will learn a mapping from this 5D input to a color and volume density corresponding to each camera ray. Then using traditional rendering methods, such color and volume density map can be rendered into a 2D image, and the loss is calculated between the rendered image and ground truth image.

The standard input to a NeRF model are images taken by calibrated cameras, where entries in the extrinsic camera matrix are known, as well as camera intrinsic parameters such as depth of field, focal length, and near and far clipping planes. However, these parameters might not always be explicitly given for every dataset. An alternative approach employed by [17] is to use COLMAP [24], which is a general purpose image based 3D reconstruction pipeline, to run structure from motion and calculate the camera pose parameters required by the NeRF model input. The input to the structure from motion pipeline can be un-calibrated camera frames from different angles for the same scene. The additional structure from motion pipeline enables NeRF to be used on a variety of different datasets, including custom dataset accessible by taking continuous frames of a scene from a smart phone. In our project, because of limited camera information of our dataset, we cannot calculate needed camera information of Nerf. Thus, we adapt the method using COLMAP to calculate necessary parameters required by NeRF. The general framework of NeRF is shown in figure 2.

3.3. NerfingMVS

NerfingMVS [30] is a recent work that focus on reconstructing 3D scenes and depth estimation. This work is based on recently proposed neural radiance fields using ScanNet dataset. In this paper, it proposes the scene-specific prior guided adaption method, which can significantly improve the depth estimation quality. It uses depth information from monocular neural network as the prior, as figure 3 illustrates. This method can also be applied on view synthesis task and the proposed guided optimization scheme is beneficial to the view synthesis quality of NeRF. The input of NerfingMVS is 2D images from different views and supervised depth information. The main idea it to guide the NeRF sampling process with an adapted depth priors from the monocular depth network.

To be more specific, NerfingMVS adds a additonal scale-invariant loss to train the depth network, which is written as follows:

$$L(D_p^i, D_{Sparse}^i) = \frac{1}{n}\sum_{j=1}^{n}|logD_p^i(j) - logD_{Sparse}^i(j) + \alpha(D_p^i, D_{Sparse}^i)| \tag{2}$$

where D_i^p is the predicted depth map and D_{Sparse}^i is the sparse depths acquired by COLMAP. The scale factor $\alpha(D_p^i, D_{Sparse}^i)$ is computed as follow:

$$\alpha(D_p^i, D_{Sparse}^i) = \frac{1}{n}\sum_{j}(logD_p^i(j) - logD_{Sparse}^i(j)) \tag{3}$$

This paper first indicates that the ScanNet dataset is suitable for view synthesis tasks, because it contains relatively abundant views and label information. Secondly, it also introduce a new Nerf-based method which improved the effectiveness of Nerf. Beside that, the paper also illustrate that NerfingMVS training time is remarkably shorter than NeRF. So in our project, we use NerfingMVS as an alternative method of NeRF, and also try to explore the effectiveness of NerfingMVS on different kind of dataset.

4. Experiment

4.1. Experiment Setup

The dataset we are using consists of 40-60 2D image inputs from different camera angles on the same indoor scene. We trained the model on a total of six scenes, four common indoor from the Hypersim [23] dataset - bedroom, office, staircase and kitchen - one bathroom scene from ScanNet, using only the RBG image data, and one custom scene we recorded using an iPhone in the Stanford Gates Computer Science building. Each scene is trained by NeRF and NerfingMVS respectively, using a single Nvidia Tesla K80 GPU over 200k iterations. Using this setup, our custom scene at

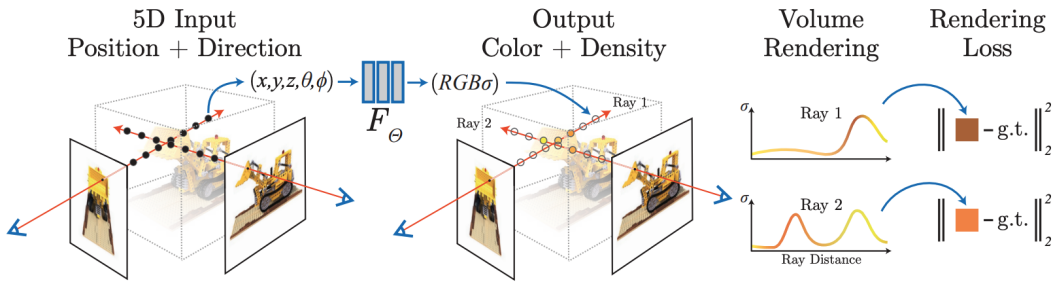

Figure 2. Framework of Nerf, cited from [18]

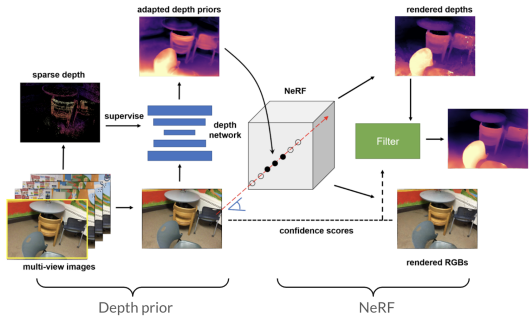

Figure 3. Framework of NerfingMVS, cited from [30]

the Gates building - which contains larger image inputs and is hence slower to train - can be trained after 35.8 hours using NeRF, and only 13 hours using NerfingMVS.

We use a structure from motion (SFM) pipeline with COLMAP to first automatically estimate the camera poses. This pipeline enabled NeRF to obtain the 5D input with only 2D image data. We implemented some adaptation had to be done on the structure from motion pipeline, because certain camera angles were far too different from the others that SFM cannot estimate its position, and we made the adaptation to automatically filter out these inputs instead of having the program crash on bad inputs.

We used the default hyperparameters to train the NeRF model as given in the configuration file for the llff_fern dataset except for two arguments - 1. we changed factor from 8 to 2 in order to get a higher resolution generation; 2. we set no_ndc to true, signaling the model that scene is not forward facing. For each scene, the first of every 8 image is taken to the test set (i.e. index 0, 8, 16, ...) and the rest is used in the training set.

4.2. Dataset

4.2.1 View Synthesis on ScanNet Dataset

The ScanNet Dataset is the dataset that the NerfingMVS model uses to compare their performance against Nerf. Therefore, to validate the reported performance, we test on one scene of ScanNet (scene0000_01).

ScanNet[6]: ScanNet is an RGB-D video dataset containing 2.5 million views in more than 1500 scans, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentations. ScanNet contains a diverse set of spaces ranging from small (e.g., bathrooms, closets, utility rooms) to large (e.g., apartments, classrooms, and libraries). Therefore, this dataset could be used for depth estimation and indoor view synthesis.

4.2.2 View Synthesis on HyperSim Dataset

Artificially synthetic dataset of multi-view scenes usually have relatively rich information and varibale views. To make systematical test on different methods, we choose a synthetic dataset to compare selected methods.

HyperSim[23]: the HyperSim Dataset is developed based on a large repository of synthetic scenes created by professional artists. It is composed of 77,400 images of 461 indoor scenes with detailed per-pixel labels and corresponding ground truth geometry as well as complete camera information, material information, and lighting information for every scene. The images in the dataset are of high quality and rich details. For example, the images from a kitchen contains a lot of dinnerware and cooking utensils on the table. This feature requires the method need to have the ability of reconstructing the fine-grained details in a scene. Beside that, the images in HyperSim dataset are from different views of a single indoor scene, and the camera position changes a lot. The changing camera position requires view-synthesis method to have robustness and high generalize ability.

4.2.3 View Synthesis on Real-World Dataset

Our custom scene taken at the Gates building consists of 58 images sampled from a video of the scene taken by an iPhone. After our automatic filtering of bad inputs in the SFM pipeline, we are left with 57 images of large varying camera angles. Some examples of the scene are shown in figure 4.

Figure 4. Custom Gates sofa scene

4.3. Evaluation

4.3.1 Evaluation Metric

In the area of view synthesis, two quantitative metrics are widely used to evaluate the model's effectiveness: peak-signal-to-ratio (PSNR) and structural similarity (SSIM). The PSNR is an expression for the ratio between the maximum possible value (groud-truth value) of a signal and the power of distorting noise (the difference between prediction and ground-truth) that affects the quality of its representation. The SSIM index aims for measure the method's effectiveness in extracting structural information from a scene. For these two metrics, higher value indicates better performance.

$$PSNR = 20 * log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \tag{4}$$

$$MSE = \frac{1}{mn}\sum_{0}^{m-1}\sum_{0}^{n-1}||f(i,j) - g(i,j)||^2 \tag{5}$$

Besides quantitative analysis, another important evaluation method of view synthesis tasks is visualization. By comparing test-set views of scenes, we can exam whether the method is able to recover the fine detail, or whether the method will generate imperfection like blurry, distortion, or ghosting artifacts.

4.3.2 Quantitative Result

The result with different view synthesis method on each dataset is shown as table 1. From the table, we can get servel important result. Firstly, NeRF performs better at most scene, especially on complex scene like HyperSim dataset. For view synthesis on indoor scene, NeRF is a greate method which can reconstruct a lot of detailed information. Secondly, the NerfingMVS method only performs better on ScanNet dataset. This might because this method mainly focus on depth estimation, and its depth prior mainly focus on optimizing depth estimation. So when the camera position and view angle changes a lot, which also indicates the depth in the image changes a lot, NerfingMVS performs bad and even can not reconstruct the scene. Finally, NerfingMVS performs better on relatively simple scene and similar camera position. The ScanNet dataset is much simpler and focus only on a small area, very few objects and narrow camera range. And our customized dataset is also simpler than synthetic dataset, which helps NerfingMVS perfom relatively well.

4.3.3 Qualitative Result

The visualization is as figure 5. We could see that among these 6 different scenes. Nerf model generates very high-quality synthesized view in most cases. Some artifacts we could see are the blurriness around the boundaries and corners. It only performs slightly worse than NerfingMVS on ScanNet scene when it generates a more vague floor. On the other hand, NerfingMVS model fails to generate a clear synthesized view on most scenes, especially on scenes with multiple objects and complicated structures. For example, all the four scenes in Hypersim Dataset have worse NerfingMVS performance than in the two real-world scenes. This could be the reason that NerfingMVS use depth estimation to guide the optimization and the depth estimation is not that accurate when the scene is too complicated. This illustrates one limitation of NerfingMVS model.

5. Conclusion

Through our experiment results, we found that NeRF generally works very well for novel view synthesis on indoor scenes. In particular, The scenes where the center objects are captured by a lot of input images are generally reconstructed nicely in a new angle.

The test time performance of NerfingMVS is not comparable to that of the original NeRF for the 4 Hypersim scenes. However, performance is comparable with faster convergence speed on scenes with simpler geometry, such as the ScanNet scene and our custom scene in the Gates building. We noticed that NerfingMVS performs much better during training time compared to test time, where both the qualitative and quantitative results are better for angles that have appeared in the training set - with the original NeRF, the difference is not that apparent. This is a sign of poor generalizability with NerfingMVS on more complex scenes.

Future work can be built on top of our findings. First, deeper diagnosis can be done to analyse the limitations of NerfingMVS's generalizability on more complex indoor geometries. Second, to address the issue that the center part of

Dataset	ScanNet-Scene0000		Hypersim-bedroom		Hypersim-office	
Metric	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NeRF[18]	21.78	**0.862**	**27.38**	**0.950**	**30.10**	**0.956**
NerfingMVS[30]	**22.60**	0.856	12.98	0.674	17.64	0.752
Dataset	HyperSim-stair		HyperSim-kitchen		Gates-sofa	
Metric	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NeRF[18]	**23.59**	**0.883**	**26.30**	**0.899**	**28.06**	**0.903**
NerfingMVS[30]	13.43	0.536	17.51	0.708	22.64	0.876

Table 1. Quantitative Result on different dataset, bold number mains better result

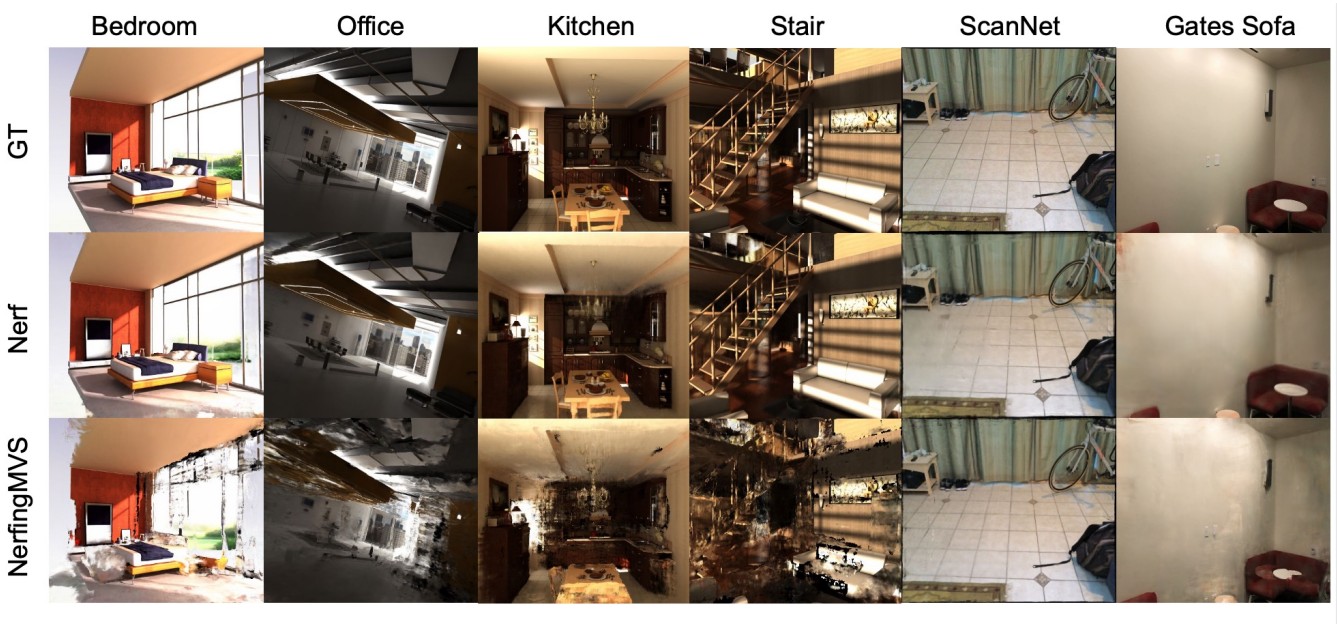

Figure 5. Qualitative Results of Ground Truth Image, Nerf and NerfingMVS on six different scenes

the scene is synthesized better than the edges, one can develop a more location specific sampling method to sample for points, so that more points on the edges can be covered in the training set.

References

[1] M. Arikan, R. Preiner, C. Scheiblauer, S. Jeschke, and M. Wimmer. Large-scale point-cloud visualization through localized textured surface reconstruction. *IEEE transactions on visualization and computer graphics*, 20(9):1280–1292, 2014. 2

[2] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. 2

[3] R. O. Cayon, A. Djelouah, and G. Drettakis. A bayesian approach for selective image-based rendering using superpixels. In *2015 International Conference on 3D Vision*, pages 469–477. IEEE, 2015. 2

[4] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013. 2

[5] L. T. B. Color. Large-scale texturing of 3d reconstructions m. *Waechter et al*, 2014. 2

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 4

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CoRR*, abs/1702.04405, 2017. 1

[8] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996. 2

[9] O. Faugeras and R. Keriven. *Variational principles, surface evolution, PDE's, level set methods and the stereo problem.*

IEEE, 2002. 1

[10] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser. Local deep implicit functions for 3d shape, 2020. 2

[11] M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klowsky, D. Steedly, and R. Szeliski. Ambient point clouds for view interpolation. In *ACM SIGGRAPH 2010 papers*, pages 1–6. 2010. 2

[12] Y. Hou, J. Kannala, and A. Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019. 1

[13] C. M. Jiang, A. Sud, A. Makadia, J. Huang, M. Niessner, and T. Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[14] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000. 1

[15] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *CoRR*, abs/2008.02268, 2020. 2

[16] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2019. 2

[17] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. 3

[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, 2020. 1, 2, 4, 6

[19] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[20] M. Niemeyer, L. M. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3501–3512, 2020. 2

[21] J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 2

[22] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Deformable neural radiance fields. *CoRR*, abs/2011.12948, 2020. 1

[23] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10912–10922, 2021. 1, 3, 4

[24] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[25] S. M. Seitz and C. R. Dyer. View morphing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996. 1

[26] S. M. Seitz and C. R. Dyer. View morphing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996. 2

[27] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. *Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2

[28] G. Vogiatzis, P. H. Torr, and R. Cipolla. Multi-view stereo via volumetric graph-cuts. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 391–398. IEEE, 2005. 1

[29] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. *CoRR*, abs/2109.01129, 2021. 1, 2

[30] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 1, 3, 4, 6