Learning Accurate and Parsimonious Point Cloud Representations from Images

Wonseok Oh
Electrical and Computer Engineering
University of Michigan, Ann Arbor
okong@umich.edu

Abstract

Methods like NeRF Mildenhall et al. (2021) that utilize volumetric neural rendering can generate top-quality view synthesis results. However, this method is optimized on a per-scene basis, making reconstruction quite time-consuming. Alternatively, deep multi-view stereo methods offer quicker scene geometry reconstruction by directly inferring the network. Using point cloud Guo et al. (2020) with the NeRF takes the best aspects of these two methods, utilizing neural 3D point clouds as well as neural features to model a radiance field. This method improves efficiency due to its ability to gather neural point features near scene surfaces using a ray marching-based rendering pipeline. This can be initialized through direct network inference, creating a neural point cloud that can be fine-tuned to exceed the visual quality of NeRF, all while benefiting from a training time. This method can also be incorporated with other 3D reconstruction methods, managing any errors or outliers through an innovative pruning and growing mechanism. Studies on several datasets including NeRF Synthetics, **DTU Jensen et al. (2014), ScanNet Dai et al. (2017) and Tanks and Temples Knapitsch et al. (2017) demonstrate that Pointcloud NeRF outperforms existing methods**, earning its place as a state-of-the-art solution.

**Bold letters haven't been done this semester. Currently working on it.*

1 Motivation

Reconstructing 3D images from 2D image data has always been a major and difficult hurdle. The fact that cave paintings from 45,000 years ago have been converted into 3D format less than a decade ago, that Monet's Cathedral series has been reconstructed into a true 3D appearance using computer graphics, and the evolution of Google Earth since 2020 shows how difficult this challenge is and only recently has it begun to make progress 1.

Present methodologies, such as Neural Radiance Fields (NeRF) Mildenhall et al. (2021) and its derivatives, despite their progress, face numerous complications. Specifically, these techniques utilize global Multilayer Perceptrons (MLPs) for reconstructing radiance fields employing a complex process of ray marching. This procedure often results in extended reconstruction periods due to the per-scene network's fitting times and the unnecessary sampling of extensive vacant spaces, negatively impacting speed and efficiency.

To address these issues, we intend to introduce an innovative point-based radiance field representation, Pointcloud NeRF, which employs 3D neural points for continuous volumetric radiance field modeling. Unlike the traditional NeRF method, Pointcloud NeRF is innovatively designed for efficient initiation using a feed-forward deep neural network Bebis and Georgiopoulos (1994) pre-trained across multiple scenes.

The project's challenges are considerable. Sophisticated data processing and computational abilities are pertinent to these methods, required to create a continuous radiance field to render high-quality

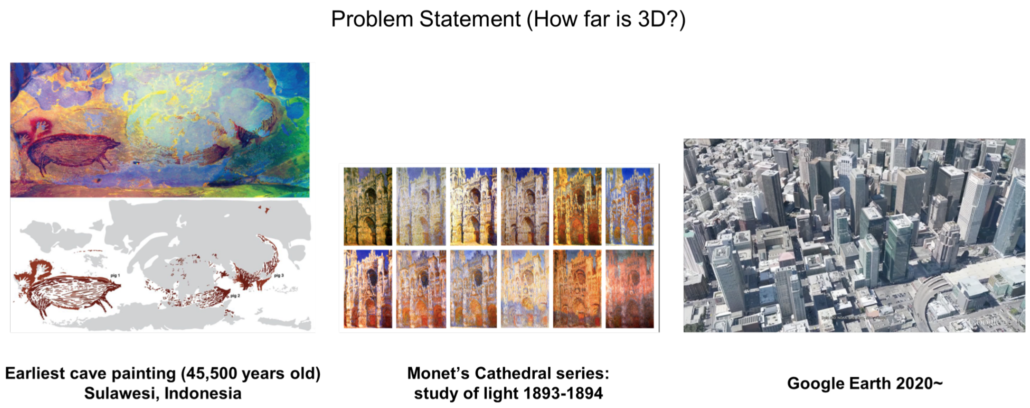
Problem Statement (How far is 3D?)
Earliest cave painting (45,500 years old)
Sulawesi, Indonesia
Monet's Cathedral series:
study of light 1893-1894
Google Earth 2020~

Figure 1: Effort to reconstruct 2D image into 3D

images. Striking a balance between maintaining quality and efficiency without compromising either is an ongoing challenge. Therefore, our pursuit of a solution that can successfully respond to these difficulties marks the non-triviality and significance of this issue.

In short, Pointcloud NeRF offers a promising direction toward effective reconstruction and photo-realistic rendering, leveraging a novel point-based radiance field representation. The learnings and advancements from our project hold great potential for propelling further breakthroughs and improvements in computer vision and graphics.

2 Related Works

2.1 Neural Scene Representation

The use of neural networks for generating realistic novel views of scenes has sparked researchers' interest in recent years. In the attempts to create more precise and authentic representations, different approaches have emerged. These include the utilization of explicit representations like meshes Yang et al. (2022), multi-plane images Flynn et al. (2019); Srinivasan et al. (2019); Mildenhall et al. (2019), and point clouds Lassner and Zollhofer (2021); Xu et al. (2022). Each of these approaches has provided strides toward developing more intricate and accurate scene reproduction.

Neural Radiance Field presents a unique approach that represents the scene as a differentiable density field. Despite its novelty and effectiveness, NeRF grapples with slow sampling speed because of volumetric ray marching. This challenge becomes a hindrance in situations that require quick sampling and rendering.

To ameliorate these limitations, subsequent works have turned their focus on optimizing training and seizing better rendering speed through space discretization Takikawa et al. (2021). Simultaneously, storage optimizations have also been analyzed as a way to enhance performance Fridovich-Keil et al. (2022). However, these solutions continue to lean on volumetric representations. While volumetric representations contribute to image accuracy, they also come with the drawback of cubic growth in encoded information, increasing the demand for computational resources.

Our proposed method deviates from these traditional methods by leveraging surface representations which offer both efficiency and parsimony. This approach addresses the formidable challenges of slow sampling speed and the high computational requirements associated with volumetric representations.

2.2 Point-based scene Representation

Recently, there have been advancements in integrating point clouds into NeRF models through the use of advanced encoders Xu et al. (2022) or parsimonious point sets Zhang et al. (2023). This novel approach includes the implementation of attraction features in the encoding process to further boost the model's performance. Our current project aims to surpass these existing models by innovatively separating 2D and 3D feature extraction in the encoding process, supported by recent research studies.

3 Method

3.1 Datasets

NeRF Synthetic datasets The base dataset is theNeRF Synthetic datasets Mildenhall et al. (2021) for our study. These data, as the name implies, are synthetic or artificially created datasets used for training or validating the NeRF models. These datasets are pivotal in posing problems and testing the performance of the algorithms [1]. This data can be easily downloaded through the URL on the official site of Matthew Tancik, author of the paper NeRF.

Neural Radiance Fields (NeRF) is a machine learning approach for 3D scene reconstruction. NeRF defines a fully connected neural network that learns to map 5D coordinates – 3D for spatial location and 2D for viewing direction – to color and opacity, effectively learning a representation for the scene's volumetric appearance.

Here, the Synthetic datasets comprise 3D scenes of varying complexity. Each scene typically consists of images captured from various positions and angles, mimicking different viewpoints. It can be used to test various aspects of 3D scene reconstruction, such as how the algorithm performs under conditions where objects' shape, texture, illumination, and other factors change.

Each scene in the Synthetic datasets is associated with RGB images, usually captured from real camera positions. In some datasets, synthetic images are created using digital 3D models, the quality of which can be compared to real 3D capture data in experiments, thus providing insights into how the NeRF model functions under a variety of circumstances.

The Synthetic datasets further aid in modeling complex physical properties such as illuminance, dispersion, and translucency, thereby facilitating more intricate 3D reconstructions.

**Currently, I used only a chair Synthetic dataset. This will be extended into various materials in the Synthetic dataset.*

Other useful datasets There are various datasets that provides not only image, but also additional information such as depth, viewing directions, and colors. The NeRF real datasets Mildenhall et al. (2021): Provides real-world images, captured from different viewpoints. Tanks and Temples DatasetKnapitsch et al. (2017): Tanks and Temples is a dataset specifically designed for benchmarking 3D reconstruction algorithms, offering detailed 3D point clouds instead of 2D images. These are datasets that could also help to create the 3D representation of the scene.

**I'm planning to use those datasets after various experiments using the NeRF Synthetic dataset*

3.2 Model Architecture

Our model is structurally bifurcated into two primary corridors: the Intermediate Neural Point Cloud generation, and the Rendering and Optimization.

3.2.1 Intermediate Neural point cloud

In the Intermediate Neural Point Cloud phase, the model recruits multi-view images to conceive an exhaustive depth for each perspective. This is achieved by harnessing the proficiency of three-dimensional Convolutional Neural Networks (G_p), exercising a cost-volume approach. Concurrently, two-dimensional traits are extracted from the assigned images by relying on a designated 2D CNN, G_f. Subsequent to the aggregation of the depth map, the model kindles a radiance field built upon these calculated points. Each point in the field is unambiguously denoted by its spatial position (p_i), an associated confidence coefficient (α_i), and image features, f_i, that remain unprotected.

3D CNNs To generate 3D point locations using cost volume-based 3D CNNs Cheng et al. (2020), the method of MVSNet Huang et al. (2018) is used. These networks produce high-quality dense geometry and generalize well across domains. For each input image I_q with camera parameters $\mathbf{p}_q$ at viewpoint q. First, construct a plane-swept cost volume by warping 2D image features from neighboring viewpoints and then regress depth probability volume using deep 3D CNNs. A depth map is computed by linearly combining per-plane depth values weighted by probabilities. Here,

[1] https://www.matthewtancik.com/nerf

Rendering and Optimization

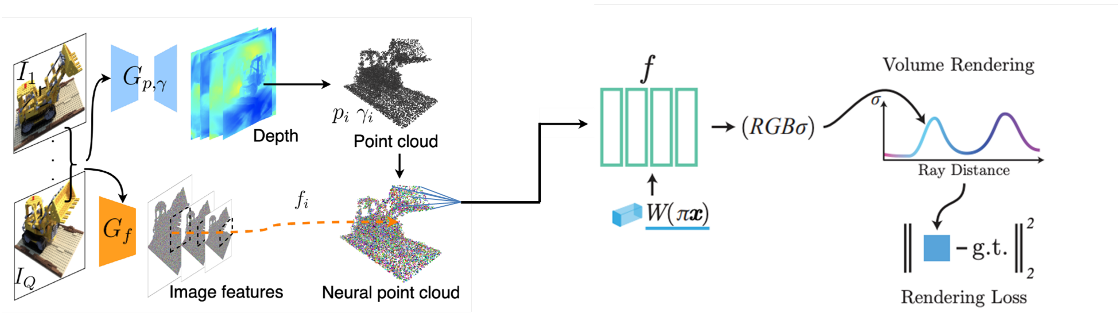

Figure 2: Overall model structure

we unprojected the depth map to 3D space to get a point cloud $\{p_1, ..., p_N\}$ for each view q. This technique is first introduced in Xu et al. (2022).

The depth probabilities describe the likelihood of the point being on the surface, thus, we tri-linearly sample the depth probability volume to obtain the point confidence c_i at each point p_i. The process can be written as:

$$\{p_i, c_i\} = G_{pc}(I_q, \mathbf{p}_q, I_{q1}, \mathbf{p}_{q1}, I_{q2}, \mathbf{p}_{q2}, ...) \tag{1}$$

2D CNN 2D CNN G_f is used to extract neural 2D image feature maps from each image I_v. Here, v is the new viewpoint provided to the 2D CNN model. The features maps are synchronized with the point (depth) prediction from G_{pc} and are used to directly predict per-point features f_i as:

$$\{f_i\} = G_f(I_v) \tag{2}$$

We specifically use a VGG network Simonyan and Zisserman (2014) architecture as downsampling layers for G_f 2.

3.2.2 Rendering and Optimization

The second phase, rendering and optimization, involves the execution of differentiable ray marching and computing shading in the vicinity of the neural point cloud, represented as (x_a, x_b, x_c). At each specific shading location, our approach clusters features from its K closest neural point neighbors and compute the volume density, ρ, and radiance, r. The latter is sequentially accumulated by capitalizing on the volume density, ρ. The outlined process is seamlessly trainable from start to finish, and the point-based radiance field can be gradually refined, aligning with the rendering loss.

Optimization Idea The optimization task is quite similar among the NeRF-based models. Our primary goal during the training phase is to minimize the distance between the rendered image I_q from its corresponding ground truth image I_{gt}. This distance metric is calculated as a weighted sum of the Mean Squared Error (MSE) and the LPIPS metric Zhang et al. (2018). Therefore, the total loss function L_{total} can be expressed as follows:

$$\mathcal{L}_{total} = D(I_q, I_{gt}) = ||I_q - I_{gt}||_2 + \lambda \cdot ||F(I_q) - F(I_{gt})||_2 \tag{3}$$

Here, F($\cdot$) denotes the perceptual feature extractor. We set the hyperparameter λ to 0.1. The Adam optimizer Kingma and Ba (2014) is used for the training of our model, which is conducted on the GreatLakes NVIDIA V100 GPU.

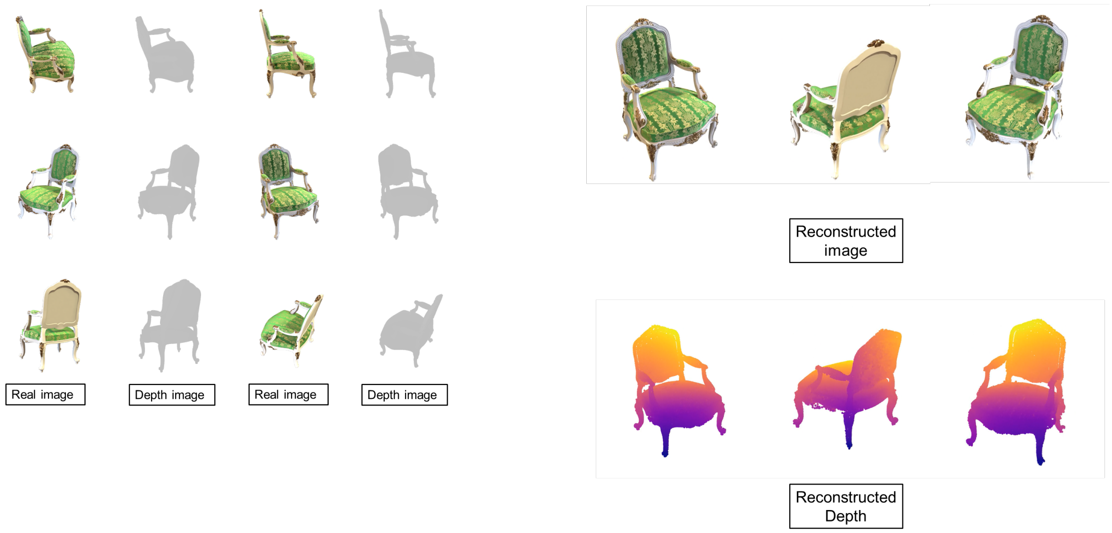

Figure 3: Qualitative result of NeRF synthetic chair dataset

4 Evaluation

4.1 Evaluation Process

In the context of this project, I initiated my approach with Point NeRF Xu et al. (2022) as the Backbone code. The project entailed training the model on a NeRF synthetic chair dataset. To augment learning capacity, modifications were made in the areas of meticulous hyperparameters fine-tuning, and refining preprocessing operations. These improvements were integrated to enhance the overall efficiency and efficacy of the learning algorithm. As mentioned in section3, the qualitative result is when the hyperparameter is 0.1.

4.2 Qualitative Results

The qualitative results 3 are based on NeRF synthetic datasets, especially the chair dataset.
**This hasn't been done this semester. Currently working on it.*

4.3 Quantitative Results

Quantitative results are PSNR, SSIM Hore and Ziou (2010), and LPIPS Zhang et al. (2018). By calculating the L_2 distance of each of PSNR, SSIM, and LPIPS for the resulting 2D images and 3D images, outputs show how similar the results are.
**This hasn't been done this semester. Currently working on it.*

5 Conclusion

The learning process for this project was an intricate one, comprising the integration of two cutting-edge theories to innovate a new concept. This was indeed challenging, with various Convolutional Neural Networks (CNN) demanding independent training. The project also entailed the application of optimization techniques learned in class, especially within the rendering process that works to convert a 2D image into a 3D one. This reiterated the fundamental fact that 3D reconstruction centers on extracting the absent data.

From a broader perspective, the project revealed inevitable strengths and weaknesses of the employed methods. While the ability to create multi-dimensional visuals stood out as a significant advantage, the complexities of multiple network training and data optimization presented some challenges. The successes of the project resonated through the successful creation of 3D representations, but the attempts also brought to light certain failures. Some methods did not prove to be as fruitful as initially presumed. There were instances when outcomes were subpar as a result of missing data and unexpected complications, underscoring the real-world complexities involved in theoretical implementations.

The most regrettable aspect was my level of capability. I realized that a considerable amount of effort from the top scholars goes into researching interesting subjects immediately. I plan to resolve this and compete with them in the future.

References

George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *Ieee Potentials*, 13(4):27–31, 1994.

Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Christoph Lassner and Michael Zollhofer. Pulsar: Efficient sphere-based neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1440–1449, 2021.

Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

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. *Communications of the ACM*, 65(1): 99–106, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019.

Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021.

Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.

Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*, pages 597–614. Springer, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Yanshu Zhang, Shichong Peng, Seyed Alireza Moazenipourasil, and Ke Li. Papr: Proximity attention point rendering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.