Machine Perception Report [Strawberry Pudding]

George Ye Kacper Ozieblowski David Sapienza Tommaso Di Mario

ABSTRACT

In this project, we tackle the problem of 3D human body reconstruction and novel view rendering using multi-view RGB images. We employ a Neural Radiance Fields (NeRF) architecture for density and RGB function approximation. Positional encodings with log-scaled Fourier features are applied to both position and direction vectors. Additionally, we explored NeuS and a geometry regularization technique inspired by RagNerf, though these did not yield satisfactory results. Our final model demonstrated improvements in both rendering quality (PSNR) and 3D surface reconstruction accuracy (Chamfer distance).

1 INTRODUCTION

Task. This project focuses on training an implicit neural network to achieve accurate 3D reconstruction and novel view rendering of human bodies from multi-view RGB images. Traditional methods often struggle with capturing fine details and realistic novel views, which are crucial for applications in virtual reality and medical imaging. Our approach aims to improve upon these limitations by leveraging the expressive power of Neural Radiance Fields (NeRF) [2], enhanced with advanced techniques for geometry representation and rendering quality.

NeRF. We are mainly inspired by the original NeRF paper and use their architecture in our final model with the exception of changing a couple of hyper-parameters. While the original NeRF paper is designed for good, realistic image rgb reconstruction and achieves good PSNR scores on its own, it struggles with accurate geometry reconstruction based on the density neural fields alone.

RegNeRF. To address this issue, we have explored two approaches, first of which is RegNerf [3]. This technique is particularly useful for sparse training datasets where the underlying geometry is ambiguous. It is well-known that real-world geometry tends to be piece-wise smooth, i.e., flat sur-faces are more likely than high-frequency structures [1]. RegNerf hinges on creating novel unseen camera poses outside the training dataset and aiming to smooth out the depth map as seen from those poses. It also uses color regularization based on normalizing flows, which we have sadly not experimented with.

NeuS. Second approach strives to improve geometry by replacing the density function with a signed distance function. The surface is then represented as the zero-level set of this SDF. To train the neural approximation of the SDF, NeuS uses a novel volume rendering framework that avoids bias during surface reconstruction.

Mip-NeRF. While the previous papers we mentioned aim to improve the geometry reconstruction, we have also attempted to use Mip-NeRF in order to reduce aliasing and blur artifacts as well as improve the ability to represent fine details. This is achieved by projecting conical frustums instead of single rays, which are in turn approximated by integrating the positional encoding over the multivariate Guassian approximate of a frustum section.

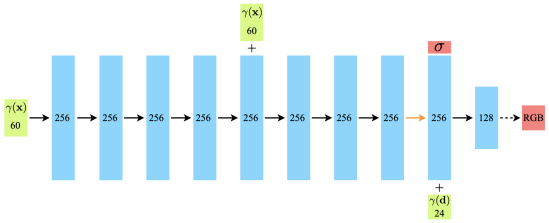

Figure 1: NeRF architecture

Final Model. Our final model, based on a NeRF architecture with dense positional encodings using log-scaled Fourier features, achieved significant advancements in both 3D mesh reconstruction and rendering from novel viewpoints. This model is evaluated based on novel view rendering performance measured by Peak Signal-to-Noise Ratio (PSNR) and 3D surface reconstruction quality assessed through Chamfer distance. Our enhancements include integrating additional layers for RGB function approximation and exploring alternative methods like NeuS [4] and geometry regularization inspired by RagNerf [3], although these did not yield optimal results within the project's timeframe.

By refining these components, our model demonstrates competitive performance improvements over the provided baseline, showcasing its potential for practical applications in digital human modeling and virtual scene generation.

2 METHOD

We use the classical NeRF approach split into the MLP color and density function approximator, the sampling and volume rendering approach and the positional encoding.

MLP approximator. To approximate the density function from the position embedding we use a simple MLP feed-forward network with 10 hidden layers of size 256 and ReLU activation functions. There is a skip connection injecting the positional embedding into the fifth layer of the network. The output is then used to predict the density which is passed through a final ReLU gate. The output feature vector is then concatenated to an embedded direction vector and passed through two additional layers of size 256 and 128 respectively before being activated by sigmoid to obtain the final color output. For additional details please refer to [2].

Positional encoding. We use the usual log-scale Fourier-features positional encoding to encode each coordinate of the position and direction vectors. We used five frequencies and five as the maximum frequency for both position and direction which amounts to vectors of size 10 for each coordinate.

Sampling. To sample points we have experimented with multi-stage hierarchical sampling using a single network akin to the one used in NeuS before reverting to the original from NeRF. The

strategy uses a coarse and a fine network. The coarse network is queried on equally spaced then randomized points along the ray, this information is then used to estimate a distribution and query new points with inverse transform sampling. These along with the original points are then provided for the fine network to train on.

Volume Rendering. Volume rendering uses exponential integration to approximate the true integral over the transmittance and density values, our approach stays identical to the one presented in NeRF.

Mesh post-processing. We have experimented with various mesh post-processing approaches. One of which was Gaussian smoothing which rendered the meshes visually appealing however it increased the Chamfer distance, we have in the end abandoned this approach. Another one we tried is the simplify quadratic decimation method inside mcubes which removes bad triangles, this increased our scores in some cases however we have not used it in the final model.

3 EVALUATION

We observed that adding an additional layer had a very big impact in geometry quality however varying other parameters like the learning rate, number of rays per image, batch size, and positional encoding frequencies hardly changed the outcomes.

4 CONCLUSION

In the limited time-frame that we spent on this project, it was very difficult to successfully implement ideas beyond the standard NeRF implementation. We have also noticed that changing the hyper-parameters can have a very big impact on performance and in particular varying the model expressiveness through the number of layers.

REFERENCES

[1] Jinggang Huang, Ann B Lee, and David Mumford. 2000. Statistics of range images. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, Vol. 1. IEEE, 324–331.

[2] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:id='cs.CV' $full_name = 'ComputerVisionandPatternRecognition' is_active = True alt_name = None in_archive = 'cs' is_general = False description = 'Cover sim age processing, computer vision, pattern recognition, and scene understanding. Roughly includes material in ACM Subject Classes I.2.10, I.4, and I.5.'$/2003.08934

[3] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2021. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. arXiv:id='cs.CV' $full_name = 'ComputerVisionandPatternRecognition' is_active = True alt_name = None in_archive = 'cs' is_general = False description = 'Cover sim age processing, computer vision, pattern recognition, and scene understanding. Roughly includes material in ACM Subject Classes I.2.10, I.4, and I.5.'$/2112.00724

[4] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2023. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. arXiv:id='cs.CV' $full_name = 'ComputerVisionandPatternRecognition' is_active = True alt_name = None in_archive = 'cs' is_general = False description = 'Cover sim age processing, computer vision, pattern recognition, and scene understanding. Roughly includes material in ACM Subject Classes I.2.10, I.4, and I.5.'$/2106.10689