Machine Perception Report [MPgroup]

Chang He Zilong Deng Kehan Wen

ABSTRACT

Differentiable neural rendering techniques have shown their capabilities to synthesize novel views and reconstruct 3D shapes. In this paper, we propose a pipeline to reconstruct the human body using NeRF given several images that are taken simultaneously from different views of a human.

1 INTRODUCTION

Given its ability to represent complicated environments, Neural Radiance Field (NeRF)[1] has been widely used in reconstructing 3D objects. However, it is not clear if NeRF can successfully reconstruct the 3D representation of a human. We attempted to reconstruct human body only given some images from different views. Our model consists of a pipeline that fine-tunes the parameters of NeRF so that both the geometry and the rendered colors are both accurate.

2 METHOD

2.1 Structure of our NeRF

The structure of our NeRF are shown in 2. Our model uses two MLPs to represent the radiance field and the volumetric density. The first MLP has 7 hidden layers of hidden size 256 and ReLU as activation functions, takes as input the positional encoded 3D coordinates, and outputs volumetric density σ and a 256-dimensional feature vector. The second MLP has 1 hidden layer of hidden size 128 and ReLU as activation functions, takes as input the feature vector and the positional encoded viewing direction, and outputs rgb color c.

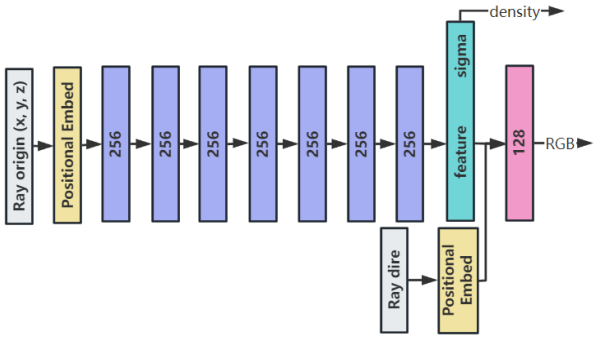

Figure 1: The structure of our nerf.

Inspired by the original NeRF[1], our pipeline utilizes positional encoding so that the model can learn from high frequencies. The format of position encoding is shown below:

$$\gamma(p) = \left(\sin(2^0\pi p), \cos(2^0\pi p), ..., \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)\right) \quad (1)$$

The number of frequencies and the maximum frequency fully characterize the positional encoding. After conducting a series of ablation studies on these two hyper-parameters for 3D coordinates and input views respectively, we found out that the optimal number of frequencies is 5 and the optimal maximum frequency is 5 for both point coordinate values and views directions.

We modified the structure of the original NeRF slightly, we replaced both the layer for density and the layer for features by one intermediate layer that output density and feature in one vector, which has a dimension of $dim_{feature} + 1$. We also apply a Softplus activation function to the density before output.

2.2 Hierarchical Sampling

The first step of our pipeline is to sample points in the 3D space. The sampling consists of two components: one is uniform sampling and the other is hierarchical sampling. First, for a given camera center and a set of ray directions, for each ray we sample points uniformly between the near boundary and the far boundary.

For the hierarchical sampling, we need to perform the volumetric rendering (mentioned in next subsection) first and obtain the coarse weight of each point along a ray, which is also the normalized cumulative density. Then we will leverage the weight to sample points. We will first calculate the CDF and get the invert CDF, which means the cumulative weight between two hierarchical points is the same. So we can sample more points in dense area to improve the accuracy of the model, see the visualization below.

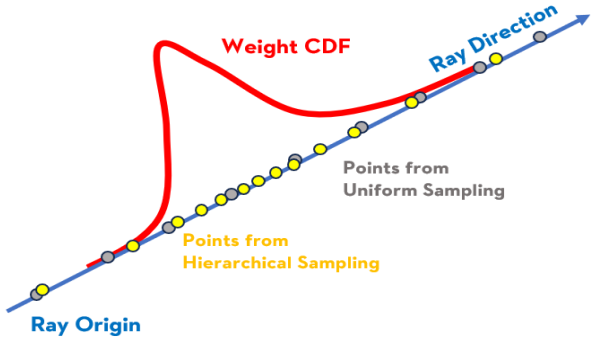

Figure 2: The structure of our nerf.

2.3 Volumetric Rendering

After calculating the RGB values and the density values of the sampled points. The third step of our pipeline is to predict the radiance field and the volumetric density for each point along the ray.

The ray color is computed using volume rendering integral 2. The transmittance T is computed using equation 3. The integration is approximated by discrete summation and their respective equations are shown in equation 4.

$$C(\vec{r}) = \int_{t_{near}}^{t_{far}} T(t)\sigma(\vec{r}_t)c(\vec{r}_t)dt \quad (2)$$

$$T(t) = \exp\left(-\int_{t_{near}}^{t} \sigma(\vec{r}_t)dt\right) \tag{3}$$

$$\hat{C} = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))c_i, T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j\delta_j\right) \tag{4}$$

3 EVALUATION

In this section, we compare experiment results between other models: template (MLP), volumetric rendering using the signed distance function (VolSDF[3]), and volume rendering using neural implicit surface (NeuS[2]). We compare models in both private and public datasets and in both MSE RGB loss and PSNR. The performance is shown in Table 1. Examples of rendered images are shown for NeRF, VolSDF, and template code. Due to space constraints, we only compare rendered images and 3D mesh from NeRF and VolSDF.

Table 1: Performance of each pipeline on each dataset

Pipeline	Public		Private	
	MSE	PSNR	MSE	PSNR
MLP	1.37e-2	27.1	-	26.9
NeRF (uniform)	3.71e-3	25.19	5.06e-3	21.78
NeRF (hierarchical)	1.36e-3	28.92	1.51e-3	28.38
VolSDF	1.5e-2	26.73	1.11e-2	23.82
NeuS	1.5e-2	26.73	-	-

Figure 3: Testing Image from NeRF

Figure 4: Testing Image from VolSDF

4 DISCUSSION

In this section, we will compare different model performances. From Table 1 we can see that our NeRF outperforms other models in rendered image quality in both public and private datasets. NeRF-based methods have demonstrated high-quality reconstruction capabilities in the human body. Compared to NeRF with only uniform sampling, NeRF with hierarchical sampling could perform better, which can be reflected in the mesh quality and the PSNR score.

On the other hand, it can be observed in Figure 5 and Figure 4 that there are undesired artifacts under the crotch of the human rendered by VolSDF. This undesired artifact is also present in the private dataset, that occurs inside the jacket. This failure could indicate that VolSDF does not work well when the ground truth image has shades so it is difficult to learn the signed distance and the radiance.

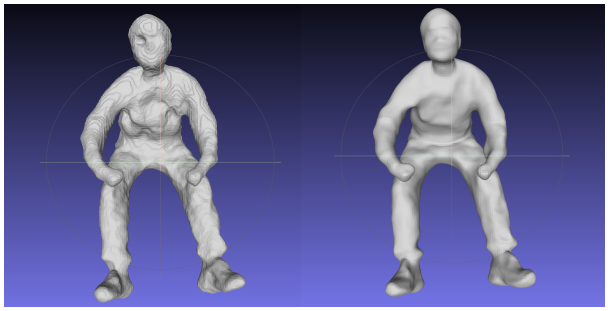

Figure 5: Mesh from NeRF

Figure 6: Mesh from VolSDF

For the mesh quality, NeRF behaves inadequately compared against VolSDF. The surface of the human is noisier and less smooth. This is due to the algorithm used in reconstructing the 3D shape. In VolSDF, the mesh is reconstructed using the learned signed distance which is optimized in the training. On the other hand, for NeRF the mesh is reconstructed using the volumetric density with a user-defined threshold. This threshold introduces biases to the reconstructed shape and the optimal threshold value is unclear. Even though VolSDF is capable to produce a better 3D mesh, its limitation in rendering shaded regions makes it our second choice to NeRF.

It is also noticeable that for different datasets, all methods perform slightly better on the public dataset. This could indicate that the human pose in the private dataset is more complicated to learn using implicit representations. Also, we use the same model architecture and the same hyperparameters to train on both datasets. This suggests that improvements can be made in fine-tuning the hyperparameters.

5 CONCLUSION

In our project, we try 3 pipelines and use the MLP pipeline as the baseline. In a word, the complicated methods like VolSDF and NeuS may produce a better mesh, but the NeRF model generally can get the highest PSNR during training. So we finally submit the models and testing results generated by NeRF pipeline.

REFERENCES

[1] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 31–46.

[2] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Conference on Neural Information Processing Systems (NIPS)*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 27171–27183.

[3] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. In *Conference on Neural Information Processing Systems (NIPS)*.