CSE 490G1 – Final Project

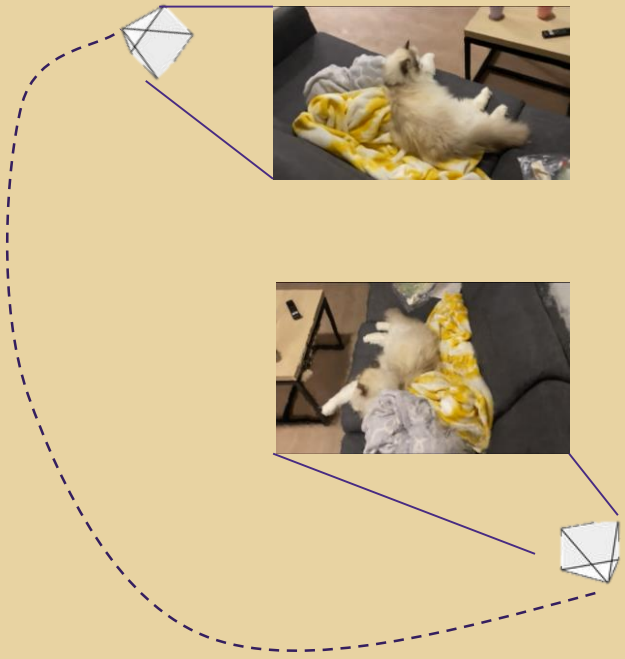
Novel View Synthesis on Real World Scenes

Jack Chuang

UNIVERSITY of WASHINGTON

PROBLEM STATEMENT

Given a video, novel views can be synthesized from unseen camera poses.



Single rgb camera path

We can obsersve a dyamic scene from a new fixed camera pose.

PROJECT SUMMARY

In my work, I adopted deep learning based methods [1] [2] to synthesize novel views on videos recorded by myself.






PROJECT SUMMARY

- Implemented NeRF [1] and D-NeRF [2] using Pytorch based on their official Pytorch implementation.
- Recorded seven videos including two static scenes and five dynamic scenes.
- These videos were processed into training data.
- Then, trained different NeRF and D-NeRF models on different scenes.

  D-NeRF: ≈ **200 hours**, ≈ **40 hours per scene** , 5 dynamic scenes, 400k iterations

  NeRF: ≈ **175 hours**, ≈ **25 hours per scene** , 5 dynamic scenes, 2 static scenes, 400k iterations

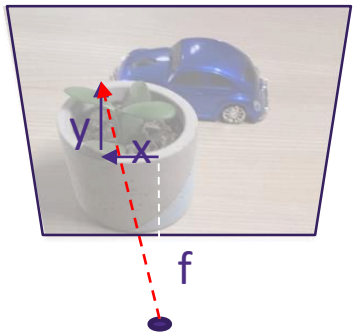
  GPU: Nvidia Tesla P100
- Quantitatively and qualitatively, compared each novel view synthesis of each scenes between its trained NeRF and D-NeRF models.

Data Collection and Processing

FFmpeg crops video into frames

Use COLMAP [3] to estimate camera translation and rotation poses in each frame and the focal length of the camera

Video recorded with a mobile phone

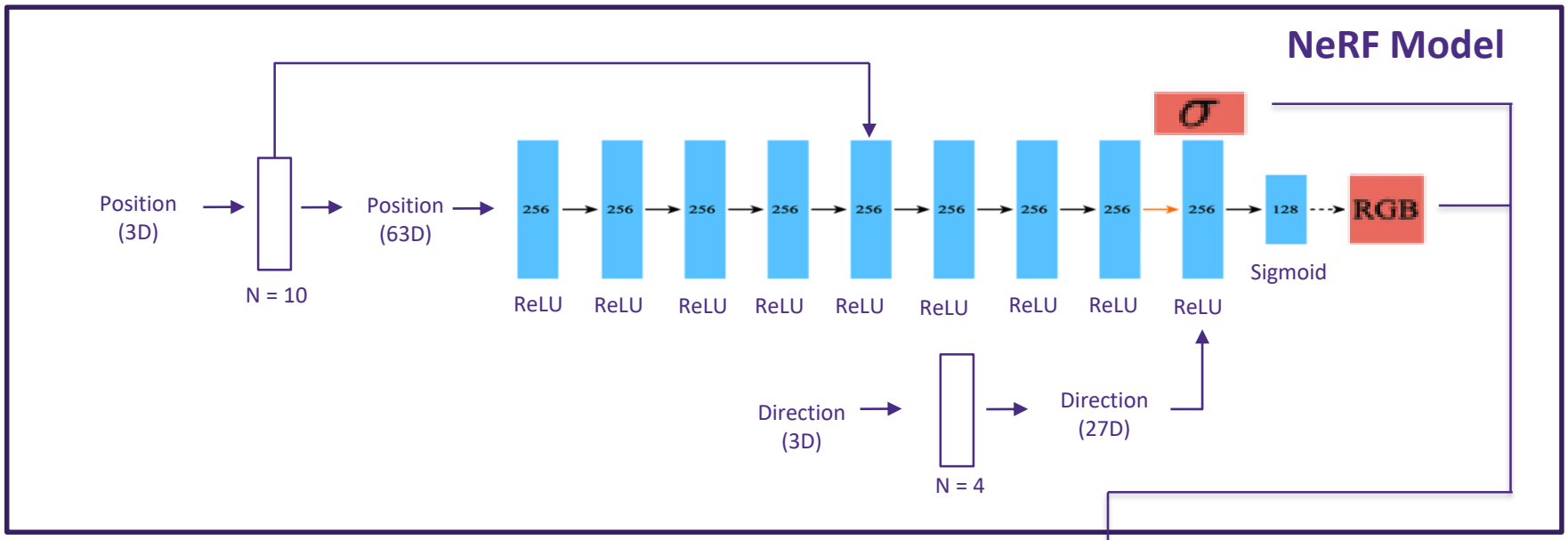


Prepare ray directions according to each camera pose and its focal length (f)

NeRF or D-NeRF

Rendering

BASELINE: NeRF

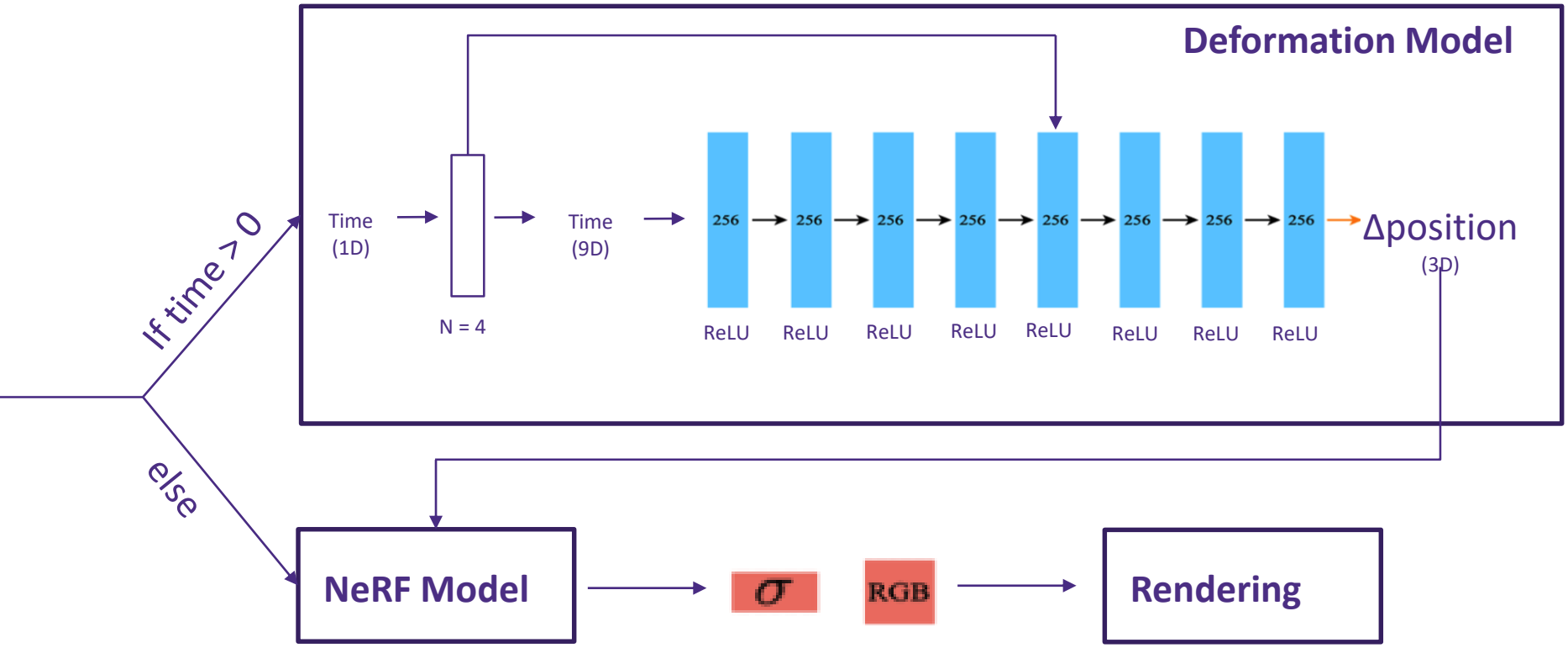


$(x) \rightarrow \boxed{} \rightarrow (x, \sin(2^k x), \cos(2^k x), \ldots)\ k = 1, \ldots, N$

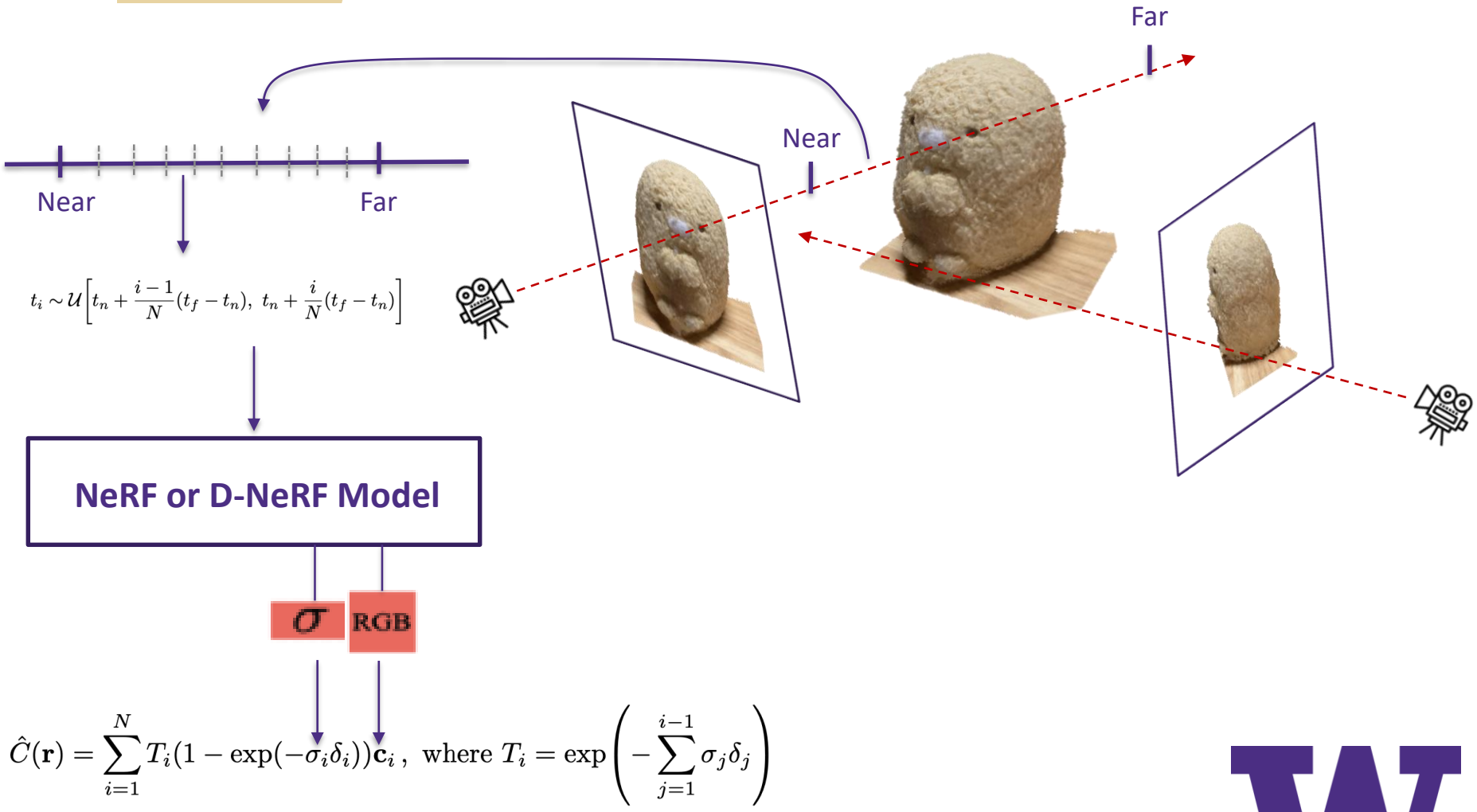
D_{in}

$D_{out} = D_{in} * (N * 2 + 1)$

Method: D-NeRF



Rendering for NeRF and D-NeRF



RESULTS

Novel View Synthesis with Fixed Camera Pose

Original Video

Synthesized (downsized 4x), 2fps

Original Video

4X speed

Camera Moves about The Target

Synthesized using validation camera poses, downsized 2x

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$
$$= 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$
$$= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE)$$

Higher the better !

Fixed Camera Pose

Synthesized using a fixed camera pose, downsized 2x

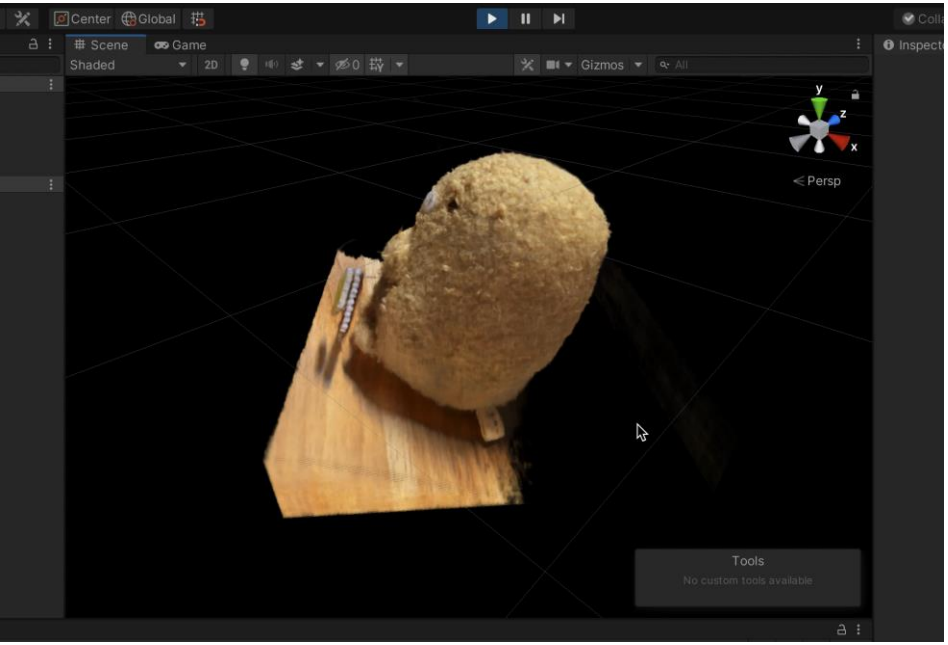
Face details can be learned by D-NeRF

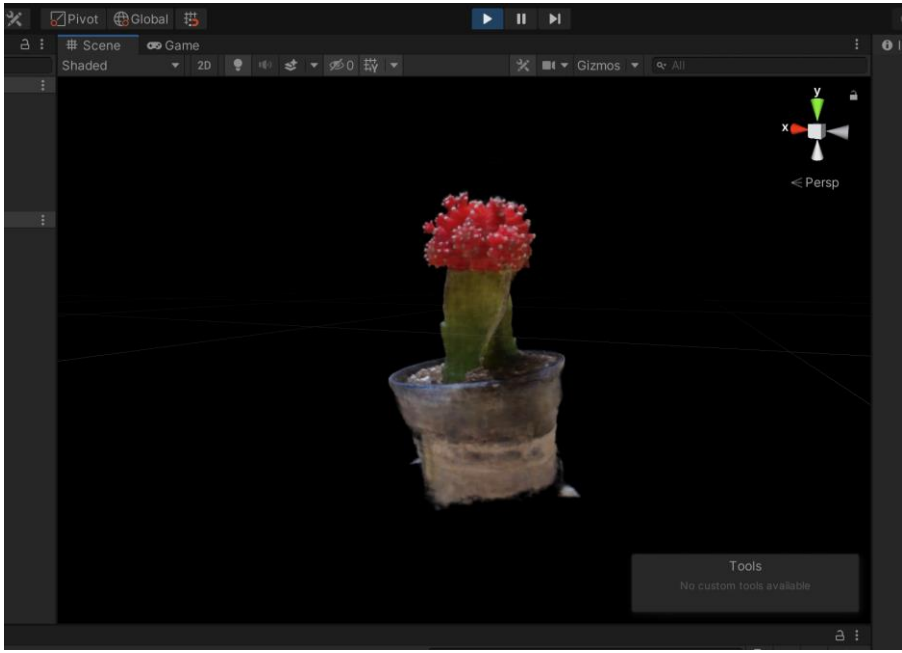
Synthesized using validation camera poses, downsized 2x

CONCLUSION

- Implemented NeRF and D-NeRF using Pytorch.
- Trained 5 NeRF and D-NeRF models for 5 dynamic scenes and compared PSNR between NeRF and D-NeRF on validation camera poses.
- The comparison shows that D-NeRF has better results than NeRF in learning dynamic scene representation.
- Given a video recorded with a moving camera, I am able to synthesize the video with a fixed camera pose.

Work in Progress





I followed [5] to render my trained NeRF models for two static scenes in Unity [4] and plan to render D-NeRF models for real world dynamic scenes in Unity.

REFERENCES

[1] Mildenhall B., Srinivasan P.P., Tancik M., Barron J.T., Ramamoorthi R., Ng R. (2020,)" NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12346. Springer, Cham. https://doi.org/10.1007/978-3-030-58452-8_24

[2] *Albert Pumarola, Enric Corona, Gerard Pons-Moll, Francesc Moreno-Noguer(2021),"* D-NeRF: Neural Radiance Fields for Dynamic Scenes "; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10318-10327 https://openaccess.thecvf.com/content/CVPR2021/html/Pumarola_D-NeRF_Neural_Radiance_Fields_for_Dynamic_Scenes_CVPR_2021_paper.html

[3] COLMAP https://colmap.github.io/

[4] Unity 3D https://unity.com

[5] https://github.com/kwea123/nerf_Unity