DLCV hw4

B10901041 林新晨

1. **(15%) Please explain:**
   a. **the NeRF idea in your own words**
      NeRF reconstructs 3D scenes from 2D images in many viewpoints. It uses neural network to encode 5D information (both position and direction information), and then it uses volume rendering to output 2D views with color and density information so that it can reconstruct realistic 3D scenes.
   b. **which part of NeRF do you think is the most important**
      I think the Hierarchical Volume Sampling part of NeRF is the most important since that is where it decides which part to focus on more for finer sampling. With the coarse network deciding which part requires more detailed sampling, the output of the fine network is then more high quality and accurate.
   c. **compare NeRF's pros/cons w.r.t. other novel view synthesis work**
      Pros: NeRF can capture nice details and realistic lighting effects that can create high quality output images.
      Cons: The training data of NeRF is difficult to prepare since it needs images from many viewpoints. Also, NeRF needs longer time to train.

2. **(15%) Describe the implementation details of your NeRF model for the given dataset. You need to explain your ideas completely.**
   I use nerf_pl from https://github.com/kwea123/nerf_pl.
   This is the parameter setting:
   --optimizer adam --num_epochs 10 --noise_std 0 --lr 5e-4 --lr_scheduler steplr --decay_step 2 4 8 --decay_gamma 0.5

My NeRF model, which is mainly nerf_pl, is an implementation of https://arxiv.org/pdf/2003.08934.pdf. It has two important parts: the positional encoding and the hierarchical volume sampling. The positional encoding uses MLP and a learned mapping function to transform input coordinates into a higher-dimensional space, which keeps a more robust representation of complex scene features. As for the hierarchical volume sampling part, it has two networks: the coarse one and the fine one. The prediction of the coarse network is used to produce more informed sampling points along each ray, so that the final output of the fine network can represent complex scenes with high-resolution details.

I also tried the ranger optimizer, and I found that both adam and ranger optimizer have quite good performance to pass the baseline after merely 2 epochs of training.

3. **(15%) Given novel view camera pose from metadata.json, your model should render novel view images. Please evaluate your generated images and ground truth images with the following three metrics (mentioned in the NeRF paper). You also need to explain the meaning of these metrics.**
   a. PSNR (Peak Signal-to-Noise Ratio): It is a ratio to quantify the relationship between the highest possible power of a signal and the power of noise that degrades its quality. For example, there are two pictures: an original picture of a TV and a copy of the picture with some quality loss. High PSNR value indicates high quality and closer fidelity to the original image. Otherwise, the picture might have some loss, like the TV might look fuzzy.
   b. SSIM (Structural Similarity Index): It evaluates the contrast, luminance as well as structural information of the two pictures, and the combination of the three values is the SSIM value. Thus, we can know how these factors collectively impact the perceived image quality.
   c. LPIPS (Learned Perceptual Image Patch Similarity): This is a way to determine the perceptual differences between two image patches using a predefined neural network. The LPIPS score is low if the result shows that a person could barely notice the changes; otherwise, the changes are obvious and the score is high.

   Different configuration settings such as MLP and embedding size, etc.

   --N_importance 64 --noise_std 0 --batch_size 1024 --lr 5e-4 --lr_scheduler steplr --decay_step 2 4 8 --decay_gamma 0.5

   I changed the MLP size:

Setting	PSNR	SSIM	LPIPS (vgg)
Optimizer ranger, num_epochs=3, D=4, W=128, skips=[2]	34.3745	0.9498	0.2317
Optimizer ranger, num_epochs=3, D=8, W=256, skips=[4]	38.5287	0.9846	0.1149
Optimizer ranger, num_epochs=3, D=12, W=512, skips=[4,8]	39.9755	0.9878	0.1065
Optimizer adam, num_epochs=10, D=8, W=256, skips=[4]	42.7558	0.9911	0.1114

Reducing the depth and width of the MLP results in worse performance, while increasing the MLP size can lead to better performance since the model can capture details better.

4. (15%) With your trained NeRF, please implement depth rendering in your own way and visualize your results.