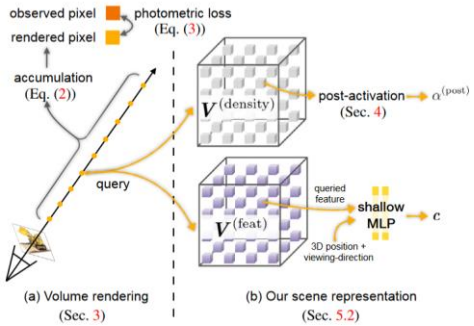
DLCV HW4 Report
r10921a36 石子仙

Problem 1

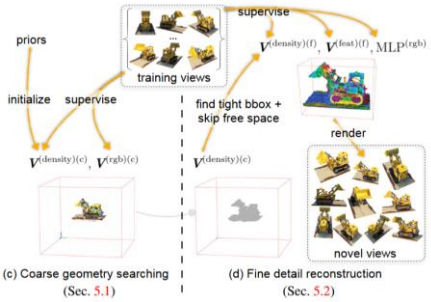
1. (5%) Please explain:
   A. the NeRF idea in your own words
      NeRF described a scene as a neural network. It considers every points on the camera ray, use their corresponding 2D view to output color and density. Then use volume rendering techniques on these values to get final RGB.
   B. which part of NeRF do you think is the most important
      I think it's the radiance field setting, it doesn't make any hypothesis on the object such as the texture, the transparency, etc. It just considers the point on the ray. I think that why it can render better quality.
   C. compare NeRF's pros/cons w.r.t. other novel view synthesis work
      The tradeoff of novel view synthesis work is time versus space. NeRF less storage(pros), less memory(pros), longer time(cons).
      NeRF can't render the new scene, it need be retrained by several viewpoints of the scene(cons).

   Please read through these two reference paper to get their ideas.

2. (10%) Describe the implementation details of Direct Voxel Grid Optimization(DVGO) for the given dataset. You need to explain DVGO's method in your own ways.
   They replace MLP, which takes large computation time, with voxel grid. To make voxel grid trainable, they use two skills. First, use post-activation. In purpose to change density to alpha value to get the sharp surface. Post-activation takes activate after doing interpolation with voxel grid. Second, they use low-density initialization. Initialize all grid values to 0, which ensure that all sampled points on rays are visible to the cameras at the beginning. For scene representation, to model volume densities, they use voxel grid method with post-activation and low-density initialization. For colors(RGB), they use voxel grid with shallow MLP.



   For scene reconstruction, they use coarse-to-fine method. First get coarse model which find a bounding box tightly, reduce the number of queried points on each ray in the later fine stage. In fine stage, they use higher-resolution density voxel grid to scaling grid.



3. (15%) Given novel view camera pose from transforms_val.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). Try to use at least two different hyperparameter settings and discuss/analyze the results.
   - You also need to explain the meaning of these metrics.
     1. **PSNR**: Peak signal-to-noise ratio (PSNR) is commonly used to quantify reconstruction quality. It's defined via MSE. The higher means the image quality is better.
     2. **SSIM**: Structural similarity (SSIM) index measures the similarity between two images. It compares luminance, contrast and structure to score and it range from -1 (not similar) to 1 (very similar).
     3. **LPIPS**: Learned Perceptual Image Patch Similarity (LPIPS) try to use neural network to learn human eyes' perception. The model will use the pretrained model to evaluate the distance between images. Evaluate the distance between image patches. Higher means further/more different. Lower means more similar.

Setting	PSNR	SSIM	LPIPS
Setting 1 Default setting of github DVGO but increase number of voxels to 256^3	35.3070	0.9753	0.0384(vgg) 0.0193(alex)
Setting 2 Default setting of github DVGO but increase number of iterations to 40000, decrease step size to 0.1	35.2348	0.9746	0.0409(vgg) 0.0219(alex)

The performance of setting 1 is better than setting 2. I think it is because number of voxels has more impact to performance then just training longer with smaller step size. The model of setting 1(2.4G) is bigger than setting two(6M), which is bigger 4(256^3/160^3) times.

Problem 2

1. (10%) Describe the implementation details of your SSL method for pre-training the ResNet50 backbone. (Include but not limited to the name of the SSL method you used, data augmentation for SSL, learning rate schedule, optimizer, and batch size setting for this pre-training phase)
   I used byol for pre-training the ResNet50 backbone, the data augmentation composes Resize(128*128) and CenterCrop(128*128), optimizer and learning rate is Adam with lr=0.0003, the

batch size is 4, and I revise the resnet50 full connected layer to (2048, 65).

2. (20%) Please conduct the Image classification on Office-Home dataset as the downstream task. Also, please complete the following Table, which contains different image classification setting, and discuss/analyze the results.

Setting	**Pre-training (Mini-ImageNet)**	**Fine-tuning (Office-Home dataset)**	**Validation accuracy (Office-Home dataset)**
A	-	Train full model (backbone + classifier)	0.362
B	w/ label (TAs have provided this backbone)	Train full model (backbone + classifier)	0.320
C	w/o label (Your SSL pre-trained backbone)	Train full model (backbone + classifier)	0.473
D	w/ label (TAs have provided this backbone)	Fix the backbone. Train classifier only	0.224
E	w/o label (Your SSL pre-trained backbone)	Fix the backbone. Train classifier only	0.374

The validation accuracy of setting: C > E > B > A > D

Train classifier only will worse than train full model. And TAs' backbone performs worse than SSL because the task (train on Mini-ImageNet) transfer badly to Office-Home domain.