DLCV HW4 Report

Student ID: R10943117
Name: 陳昱仁

Problem 1.

1. **Please explain:**
   a. **the NeRF idea in your own words**
   b. **which part of NeRF do you think is the most important**
   c. **compare NeRF's pros/cons w.r.t. other novel view synthesis work**

a. NeRF 的核心概念是將物體以及場景資訊，encode 進類神經網路中，接著使用 computer graphics 中的 volume rendering 將神經網路中的資訊投影出來。之後就可以對這個物體與場景 render 出連續、不存在於原始資料的視角。

b. 用類神經網路代表光場本身，直接利用到圖學的知識，而不是像處理黑盒子般，套用沒有空間意義的 CNN，NeRF 只利用 MLP 完成整個 work。

c. Pros:
   1. 網路非常的小，只用了 5MB 的網路就可以生成 1008x752 的高解析度影像
   2. 透過積分，可以直接取得場景的 depth map
   3. 對於被描述的場景沒有任何假設，可以處理金屬和半透明物體

   Cons:
   1. 一個類神經網路只能描述一個場景
   2. 生成影像耗費時間
   3. 網路在訓練初期不容易穩定

2. **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.**

DVGO 相比 NeRF 訓練速度更快，其效能也堪比 NeRF，作者提出兩個方法使訓練速度加快以及使質量提升。第一個是引入 post-activation interpolation 在 voxel dnesity，他能夠在較低的 grid resolution 產生更清晰的表面，第二是 direct voxel density optimization 容易產生幾何次優解，使用一些 priors 鞏固訓練過程。

我的實作 detail 如下，shallow MLP layer with 128 channels，optimizer 使用 Adam，batch size 為 8,192 rays，coarse and fine iterations 分別為 20k 和 30k，base lr 0.1 for all voxel grids and 0.001 for MLP，並使用 exponential learning rate decay。

3. **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.**

PSNR: 測量圖片質量指標，衡量最大值信號和 noise 之間的參考值，單位為 dB，其值越大，圖片失真越少。

SSIM: 衡量兩張圖相似程度的指標。二圖的結構相似性可以看成是失真影像的影像品質衡量指標，數值越大越好，結構相似性相比 PSNR，在影像品質的衡量上更能符合人眼對影像品質的判斷。

LPIPS: 又稱為 perceptual loss，建立在一個人類判別的感知 dataset，利用深度學習讓 AI 學習兩圖

片間的相似度，和傳統只使用函數計算的 SSIM 和 PSNR 相比，更能反應出人類的感知情況，數值越小越好。

Setting	PSNR	SSIM	LPIPS (vgg/alex)
fine_iter = 20000 fine_voxels=160**3	35.15	0.974	0.042/0.022
fine_iter = 40000 fine_voxels=320**3	35.24	0.975	0.040/0.021

從上方結果可以看到，第二個 setting 的結果比第一個要好，和 setting 的改變相當合理，增加 finetune 的 iteration 以及增加 finetune 時 voxel 的數量都可以有效地增加效果，但是訓練的時間會比較久，即拿時間來換取效能。

Problem 2.

1. **Describe the implementation details of your SSL method for pre-training the ResNet50 backbone.**

我使用的 SSL 是 BYOL，batch 設定 16，epoch 設定 500，lr 設定 0.0003，lr 每 50 epoch 變成原本 0.8 倍，optimizer 為 Adam，data augmentation 如下圖所示。

```
transform = transforms.Compose([
    transforms.Resize(128),
    transforms.RandomHorizontalFlip() ,
    transforms.RandomAffine(10, translate = (0.1 , 0.1),
                            scale = (0.9 , 1.1)),
    transforms.ColorJitter(brightness = 0.15,
                           contrast = 0.15,
                           saturation = 0.15,
                           hue = 0.15),

    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406],
                         std=[0.229, 0.224, 0.225])
])
```

2. **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	Validation accuracy
A	36.45%
B	52.46%
C	57.14%
D	41.13%
E	44.58%

首先先看 A，A 沒有 pretrain 在任何 dataset 上，在 office dataset 很小的情況下，表現是最不好的，可見事先先 pretrain 在別的 dataset 上訓練 backbone 是很有幫助的。B 以及 C 分別為助教提供的 pretrain weight 以及我自己 train 的 pretrain weight，可以看到 C 比 B 好，代表我 pretrain 階段的設定比助教提供的佳，且其對應的 D 以及 E 也是我的比較好，然而 D 以及 E 比 B 和 C 來的差，原因來自他們 freeze backbone，因為 pretrain 的 mini-ImageNet 和 office-Home 還是存在一定的 bias，所以允許 backbone 訓練修正才會訓練的比較好，freeze 的話 backbone 只會學習來自 mini-ImageNet 的資料，自然會比在 office-Home 上有修正的 backbone 表現來的差。