DLCV hw4 電機碩一 R11921100 溫威領

**Problem 1: 3D Novel View Synthesis**

1. Please explain:

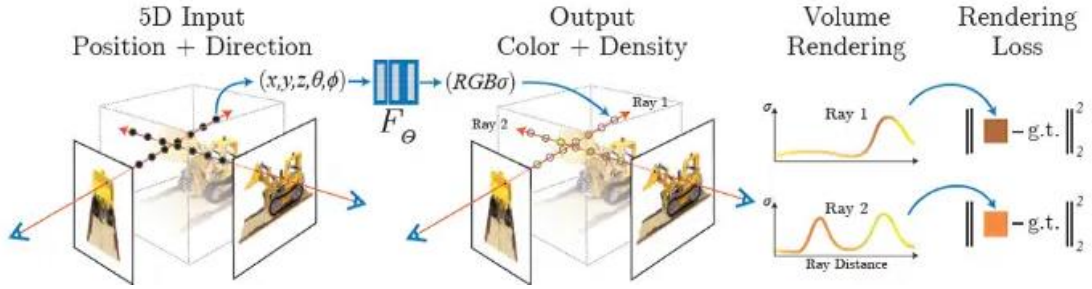
a. the NeRF idea in your own words

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 的主要想法是使用神經網路來學習 3D 場景中每個像素點的顏色和透明度信息，通過插值的方式從這些信息中合成出新的視角，並渲染出 2D 圖像。神經網路主要採用 MLP，它的輸入是一組 3D 位置(x, y, z)和觀察方向(θ, φ)，並希望輸出該觀察路徑上每個點的顏色(r, g, b)和體積密度($\sigma$)，

b. 我覺得 NeRF 最重要的部分就是下面這張圖，綜觀了整個 Nerf 的過程，但一些細節如 position encoding 也是提升精確度不可或缺的，而整體最困難的點還是在於有新的想法並實踐出來(3D 場景的表示、MLP 的輸入與輸出等)。




c. 與其他新型視覺合成技術相比，NeRF 的優點在於它可以生成更高精確的 3D 場景，並且能夠解決傳統方法中的許多問題，如陰影和物體的遮擋等。但它的缺點在於計算量較大，訓練和渲染速度很慢，需要較長的時間來訓練模型，而且在一個場景上訓練的 NeRF 模型無法用於其他場景。

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 會根據給定的數據集構建體素網格，網格中的每個體素都有一個初始的顏色和透明度。然後不斷更新體素的顏色和透明度，直到收斂為止。DVGO 採用 coarse to fine 的訓練方式。在 coarse 階段，使用先驗信息和多視角圖像訓練兩個粗粒度的 voxel，然後使用其中的密度場確定場景中的空白區域。在 fine 階段，利用 coarse 階段確定的密度場可以得到更緊密的 bbox，可以將 grid 定義在這個 bbox 內，減少無關變量的訓練。並且在體渲染的時候還可以通過粗的密度場提前得知射線上哪些空白點和被遮擋的無用點應當被跳過，所以比 NeRF 的訓練速度快上許多。

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. Try to use at least two different hyperparameter settings and discuss/analyze the results.

| **Setting** | **PSNR** | **SSIM** | **LPIPS (vgg)** |
|---|---|---|---|
| coarse_train_iters=5000<br>fine_train_iters=20000<br>coarse_num_voxels=1024000<br>fine_num_voxels=(160**3) | 35.17326808 | 0.974450365 | 0.041290114 |
| coarse_train_iters=10000<br>fine_train_iters=50000<br>coarse_num_voxels=3*1024000<br>fine_num_voxels=3*(160**3) | 35.35992661 | 0.975739255 | 0.037858232 |

PSNR(Peak Signal-to-Noise Ratio，峰值信噪比)衡量原圖和處理後圖像之間的差異程度，數值越高表示圖像品質越好。SSIM(Structural Similarity，結構相似度)衡量了原圖和處理後圖像之間的結構相似度，範圍為 0~1，越大代表圖像越相似，當兩張圖片完全一樣時 SSIM 值為 1。LPIPS(Perceptual Similarity，感知相似度)衡量了人類感知系統對圖像的辨識能力，數值越小表示越好。

從上表結果來看，num_voxels 與 train_iters 增加，三項圖像品質的指標皆有變好，有稍微提升圖片品質，但增加 num_voxels 導致 model 參數量的大量提升與品質提升幅度並不太對等。

**Problem 2: Self-Supervised Pre-training for Image Classification**

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

使用 BYOL : https://github.com/lucidrains/byol-pytorch

Backbone: Resnet50

```
backbone = models.resnet50(weights=None)
```

SSL method: SimSiam

```
learner = BYOL(backbone, image_size=128, hidden_layer='avgpool', use_momentum=False)
```

Optimizer: Adam, learning_rate=3e-4

```
optimizer = torch.optim.Adam(learner.parameters(), lr=3e-4)
```

Data augmentation: BYOL lib default

Batch size: 32

Input image size: 3*128*128

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.

Classifier:

```
classifier = nn.Sequential(nn.BatchNorm1d(1000), nn.ReLU(), nn.Dropout(p=0.5),
                           nn.Linear(1000, 512), nn.BatchNorm1d(512), nn.ReLU(), nn.Dropout(p=0.5),
                           nn.Linear(512, 65))
```

| **Setting** | **Pre-training (Mini-ImageNet)** | **Fine-tuning (Office-Home dataset)** | **Validation accuracy (Office-Home dataset)** |
|---|---|---|---|
| A | - | Train full model | 35.47% |
| B | w/ label | Train full model | 44.09% |
| C | w/o label | Train full model | 52.46% |
| D | w/ label | Fix the backbone. Train classifier only. | 22.66% |
| E | w/o label | Fix the backbone. Train classifier only. | 37.68% |

從上表五個 setting 中，model C 有最高的 validation accuracy (52.46%)，也就代表 SSL pre-training + full model fine-tuning 的效果是最好的。再從 B、C 和 D、E 兩個組合中 fine-tuning 條件相同下比較，可看出 without label pre-training 的 SSL method 皆比 with label pre-training 來的好(C > B、E > D)。而五個 model 中 setting D 是最差的，從 B、D setting 推測是 fine-tuning 時 fixed backbone 所影響，因為 model D 只有後面的 classifier 被 train，而 mini dataset 與 office dataset 的 latent space domaind 可能又有不小落差，導致其 validation accuracy 最低，甚至比沒有 pre-training 但 train full model 的 model A 還低(D < A < B)。