DLCV Fall

醫工四 郭思言 B08508002

Hw4

Part 1:

1. Explain:
   (a) The NeRF idea:
   - 這篇論文主要是希望可以利用 NeRF 的概念來達到 3d view synthesis 的目的。作者希望在針對特定場景給定幾張不同角度的影像的情況下，可以重建出這個場景的 NeRF。如此便可以利用這個 NeRF，render 出任何 viewpoint 下的影像。
   - 每一個不同角度、靜態的影像可以表示為 position(x, y, z)以及 camera 視角(θ, φ)這樣子 5d 的連續 function（為一個 5d 的 coordinate）。提供幾張不同角度的影像，為了要能夠 render 出任意視角的影像，作者希望可以利用 MLP，提供 position(x, y, z)以及 camera 視角(θ, φ)，訓練出這個空間中任意位置的 density 以及在那個視角下的 RGB。如此，便能夠利用既有的 volume rendering 的技術來得到給定視角的影像。
   - 利用 density 的 volume rendering 方法大概想法是一個視角的 ray 如果遇到 density 很大的障礙物的話，那麼看到的顏色應該會由這個 density 很大的東西所顯現，而後面的物體應該會被擋住。因此只要有了每個位置的 density，便可以 render 出任意視角看到的影像。為了可以加速也有 coarse 以及 fine 兩步驟等等許多方式可以使用。

   (b) Which part of NeRF is the most important:
   - NeRF 這篇中我認為有許多方法都非常重要，最重要的想法大概是可以利用 Multiple Layer Perceptron (MLP)，而非其他例如說 CNN 等等的架構來將 position 以及視角訓練出空間中的 RGB 以及 density。
   - 此外，我認為也很重要的是先將原來的 coordinate 進行 positional encoding，以便這個 MLP 可以學習到高頻的特徵，如此進行訓練可以提升效果。

   (c) Compare NeRF pros/cons wrt other novel view synthesis works:
   與一些作者拿來做比較的架構的差異：
   - Neural Volumes (NV): 利用的方法是 Convolutional Neural Network 來預測 RGBα，這個方法還需要額外取得背景資訊。
   - Scene Representation Network (SRN): 這個方法是與作者相同是利用 MLP 但是是將每個 xyz 預測出一個 feature vector。此外，還會再訓練一個 recurrent neural network 來 predict 下一步的 step size，在罪中的最後一部可以 decode 成那個位置單一的 RGB 顏色。

- Local Light Field Fusion (LLFF): 這篇的做法是是訓練一個 Convolutional Neural Network 來 predict RGBα，之後透過 alpha 合成和將附近的 MPI 混合到新的視點中來呈現出影像。

主要的不同大概在於利用 MLP 或是 CNN，以及希望能夠預測出的東西以及最後的成像方式。此外這些方法的主要 tradeoff 就是時間與空間，無法兼得。

Pros: NeRF 的成像效果都是較目前其他的方法還要來得好的，能夠得到更清晰的影像。NeRF 要儲存的 model weight 也是最小的，只要 5GB。

Cons: NeRF 在 inference 階段會花較多時間，而像是 LLFF 會花較少時間，僅需要 10 分鐘即可。

2. Describe the implementation details of Direct Voxel Grid Optimization (DVGO) for the given dataset (explain DVGO methods):
   - DVGO 主要是基於 NeRF 這篇 paper，希望可以提升訓練上的速度。
   - DVGO 優化 training 時間的主要方法是使用“dense voxel grid”來建立 3d 模型。在模型架構上和 NeRF 一樣是使用 MLP。直接訓練 dense voxel grid 可以達到非常快速的收斂，但是容易卡在 suboptimal 的解，因此作者提出的解決方法為在空間中分配“cloud”並使用 photometric loss。作者是說如此可以增加訓練的穩定性。
   - 在執行上，首先以非常接近 0 的透明度還初始化 dense voxel grid。再來就是作者以較低的 learning rate 進行 voxel 的訓練。
   - 此外，作者也提出了“post activation”的方法來提升 render 出的影像的品質。原來大多利用的 interpolation 的方法，使用 pre/in-activation 可能造成輸出影像較為光滑、模糊。使用 Post activation 可以相較其他方法輸出更尖銳的線性表面與邊界，可以得到更高品質的影像。
   - NeRF 以及其他更早的研究的 grid resolution 可能需要 512**3 ~ 1024**3，而 DVGO 這篇作者是提到只需要使用 160**3 的 grid resolution 就可以得到和其他差不多品質的影像了！
3. Given novel view camera poses from transforms_val.json, your model should render novel view images. Evaluate generated images and ground truth images with the following metrics:

Setting	PSNR	SSIM	LPIPS
Settings 1 - DVGO original github default settings: Num_voxels = 160^3 Step_size = 0.5 N_iters(coarse) = 20000 N_iters(fine) = 50000	35.27801747322083	0.9752489881081897	0.03961326129734516 (vgg) 0.021313915941864253 (alex)

Settings 2 - modified base on the paper’ s implementation (some are different): **Num_voxels = 320^3** **Step_size = 0.5** **N_iters(coarse) = 20000** **N_iters(fine) = 50000**	35.219830417633055	0.9747940201415999	0.0383372712880373 (vgg) 0.018747738897800445 (alex)

```
Testing psnr 35.27801747322083 (avg)
Testing ssim 0.9752489881081897 (avg)
Testing lpips (vgg) 0.03961326129734516 (avg)
Testing lpips (alex) 0.021313915941864253 (avg)
```

⇨ 調整後的 settings 2 model 變得較大，不過效果卻沒有提升，應該是 overfit 了一些。

Part 2:

1. Describe the implementation details of your SSL method for pre-training the ResNet50 backbone.
   - 我選擇使用 BYOL - LINK 來 pre-train 我的 ResNet50 backbone.
   - SSL 總共分為二階段：pretrain 以及 finetune。
   - Pre-train 是在沒有 label 的情況下進行，BYOL 使用 online 以及 target 兩個 network 來進行 training，訓練 online network 來預測 target network 在同個影像的輸出結果。BYOL 的訓練是要預測前一次的輸入 label，因此沒有使用任何的 negative pair。
   - Finetune 時，我在 pretrain 完的 ResNet50 後面拔掉原來的 fc 層後以另外三層 fc 曾取代，分別是 2048 -> 2048 -> 1000 -> 1000，最後輸出是 65 個 class。

     Hyperparameters and training details:
   - Pretrain: epoch: 250, batch size: 512, optimizer: Adam, learning rate: 3e-4, data augmentation 的部分我只使用了 random horizontal flip。
   - Finetune: epoch: 80, batch size: 64, optimizer: Adam, learning rate: 3e-4, loss function: cross entropy loss, data augmentation 我使用了 image net 的 auto augment policy，還使用了 random resized crop, color jitter, random affine 等等。（對於 A, B, C, D, E 五個不同的 settings 皆是一樣的設定）
   - （image size 都是按照規定固定到 128 x 128）
2. Please conduct image classification on Office-Home dataset as the downstream task. Complete the following table.

Setting	Pre-training (Mini-ImageNet)	Fine-tuning (Office-Home dataset)	Validation accuracy (Office-Home dataset)
A	-	Train full model (backbone + classifier)	accuracy: 0.4828 (Epoch: 75)
B	w/ label (TAs have provided this backbone)	Train full model (backbone + classifier)	accuracy: 0.4877 (Epoch: 75)
C	w/o label (Your SSL pre-trained backbone)	Train full model (backbone + classifier)	accuracy: 0.5887 (Epoch: 61)
D	w/ label (TAs have provided this backbone)	Fix the backbone. Train classifier only	accuracy: 0.0961 (Epoch: 74)
E	w/o label (Your SSL pre-trained backbone)	Fix the backbone. Train classifier only	accuracy: 0.3867 (Epoch: 80)