

# hw2\_r13921068

## DLCV HW2 Report

學號：r13921068

姓名：吳家萱

### Problem 1: Conditional Diffusion Models

1. Describe your implementation details and the difficulties you encountered.

我的實作方法參考了 [dome272/Diffusion\\_Models\\_Conditional\\_Model](#) 的 Github repo 寫法，Conditional 的方法參考了 [Classifier-Free Diffusion Guidance](#)，UNet\_conditional 寫法參考了 [Kaggle Deffusion & Conditional-Deffusion\\_Fash-MNIST](#)：

1. 新增了兩個嵌入層：label\_emb 和 dataset\_label\_emb，用於處理類別標籤 (labels) 和資料集標籤 (dataset\_labels) 。
2. 在 forward 時如果 labels 不為空，則取出對應的標籤嵌入 label\_emb，並將它加到 t 上，這樣模型在擴散時就能考慮到特定類別的圖像特徵；同樣，如果 dataset\_labels 不為空，則取出資料集標籤嵌入 dataset\_label\_emb，也加到 t 上，這樣模型也會考慮到數據集的特性，最後將這個帶有時間和標籤信息的嵌入向量 t 傳入 unet\_forwad 方法中進行圖像生成。

在兩個資料集的處理上，我的方法如下：

- 對於每個資料集，分別讀取對應的圖像檔案路徑與標籤，並增加一個 dataset label，用來區分圖像屬於 MNIST-M (標記為 0) 或 SVHN (標記為 1) 。

模型參數：

Learning Rate	Optimizer	Batch Size	Epochs	Loss
1e-3	AdamW	64	1000	nn.MSELoss()

遇到的困難：

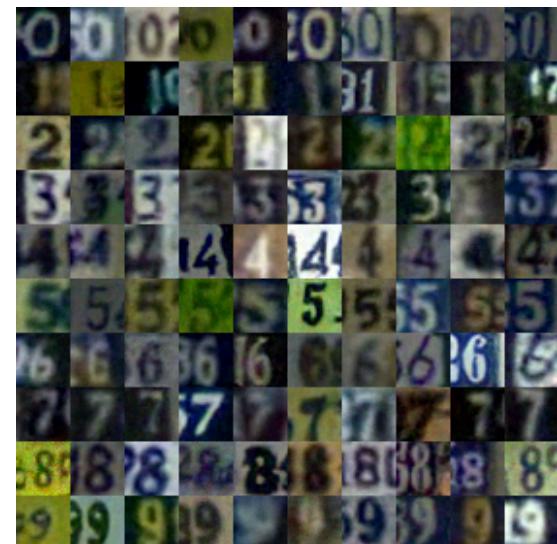
- UNet\_conditional 配合題目兩個資料集的設計，在調整上花了許多時間，尤其是 UNet\_conditional forward 的結構，和調整 Up forward 的結構。
- 一開始我 epoch 只有設大約 400 左右，但發現大約在 epoch 370 時，MNIST-M 的準確率已經可以達到 95% 以上，但 SVHN 却只有 58%，所以後來透過調整 epoch，直到 epoch 1000 SVHN 的準確率才順利達到 98%。

2. Please show 10 generated images for each digit (0-9) from both MNIST-M & SVHN dataset in your report. You can put all 100 outputs in one image with columns indicating different noise inputs and rows indicating different digits. [see the below MNIST-M example, you should visualize BOTH MNIST-M & SVHN]

- MNIST-M



- SVHN



3. Visualize a total of six images from both MNIST-M & SVHN datasets in the reverse process of the first “0” in your outputs in (2) and with different time steps. [see the MNIST-M example below, but you need to visualize BOTH MNIST-M & SVHN]

- MNIST-M

t = 0	t = 200	t = 400	t = 600	t = 800	t = 1000

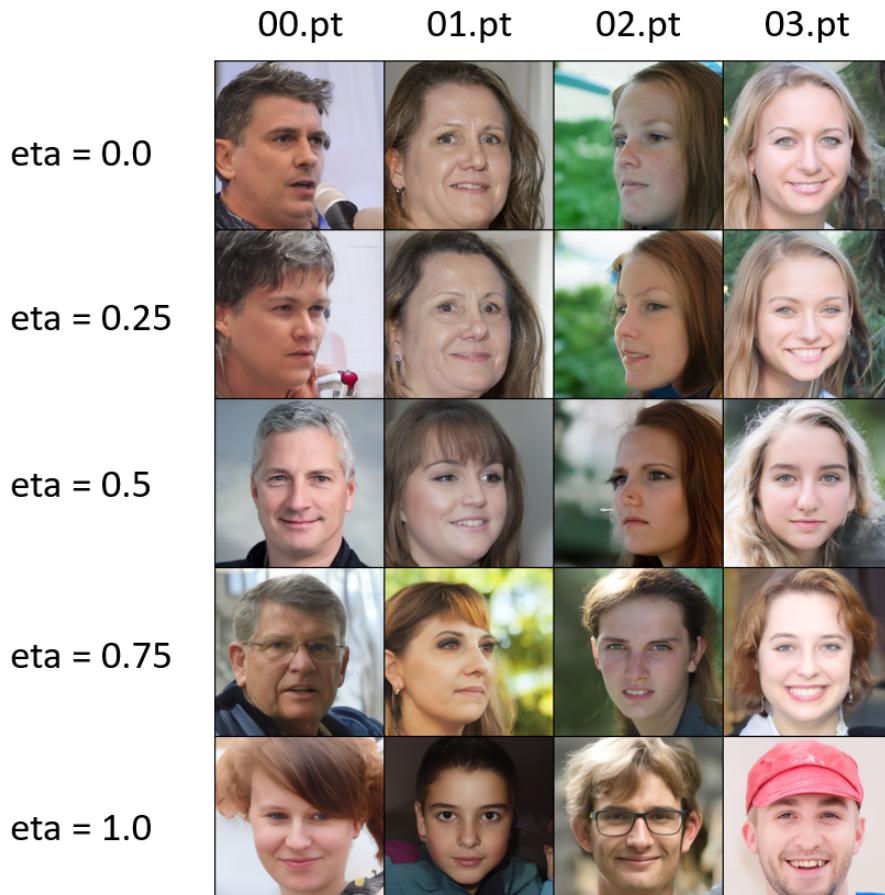
- SVHN

t = 0	t = 200	t = 400	t = 600	t = 800	t = 1000

---

## Problem 2: DDIM

1. Please generate face images of noise 00.pt ~ 03.pt with different eta in one grid. Report and explain your observation in this experiment.



- eta = 0.0 : 和 Ground Truth 十分相像，生成的圖像非常一致，沒有隨機性，代表此時產生的結果非常穩定，但缺少隨機多樣性。
- eta = 0.25 ~ 0.75 : 圖像開始顯示出多樣性，有一些臉部特徵的變化，可以觀察到當 eta 的值越大，生成過程中的隨機性就越多，也更接近 DDPM，此時產生的結果有著更多的隨機多樣性。
- eta = 1.0 : 圖像變得非常多樣化，此時的 DDIM 相當於 DDPM，每次生成的結果都會有所不同，此時產生的結果圖像具有更多的隨機性和多樣性。

2. Please generate the face images of the interpolation of noise 00.pt ~ 01.pt. The interpolation formula is spherical linear interpolation, which is also

known as slerp.

What will happen if we simply use linear interpolation? Explain and report your observation. (There should be two images in your report, one for spherical linear and the other for linear)

- **spherical linear interpolation, slerp**



- 從左到右的圖像顏色過渡比較自然，因為 slerp 使用了球面線性插值，考慮了高維空間中向量之間的角度差異，能夠更好地保持原始向量的方向一致性；也就是 slerp 在進行插值時，保持了噪音的特徵向量方向，因此生成的圖像在顏色和特徵上更加穩定。
- 圖像中的人物面部特徵隨著插值參數  $\alpha$  的變化逐漸過渡，五官的變化是連續且自然的。

- **linear interpolation**



- 可以明顯觀察到中間幾張圖像出現了綠色的顏色失真，可能是在線性插值的過程中，當兩個噪音向量之間的差異過大時，直接線性插值可能導致生成的中間向量無法很好地保持圖像的色彩一致性，特別會發生在高維空間的插值過程中。
- 另外，中間幾張圖像明顯變得模糊，特徵逐漸消失，幾乎看不清楚，可能是線性插值在中間步驟時，插值向量偏離了數據的有效範圍，導致生成的圖像失去了原有的清晰度和特徵，導致圖像失真。

---

### Problem 3: Personalization

1. Conduct the CLIP-based zero shot classification on the hw2\_data/clip\_zeroshot/val, explain how CLIP do this, report the accuracy and 5 successful/failed cases.
- 我參考 [CLIP Github repo](#) 提供的 Zero-Shot Prediction Example Code，使用 ViT-B/32 模型進行撰寫。

- CLIP 模型進行推理時，沒有預先針對特定類別進行訓練，而是會同時對圖像和文本進行編碼，然後根據這些編碼來計算圖像與文本之間的相似度，這使CLIP 能夠在沒見過的新類別上進行分類，即使在訓練過程中沒有見過這些特定的標籤。
- Accuracy: 58.68%
- 以下皆為 label "7": "fox"：
  - 5 successful cases:

Image	7_488	7_489	7_490	7_491	7_492
Right label	fox	fox	fox	fox	fox

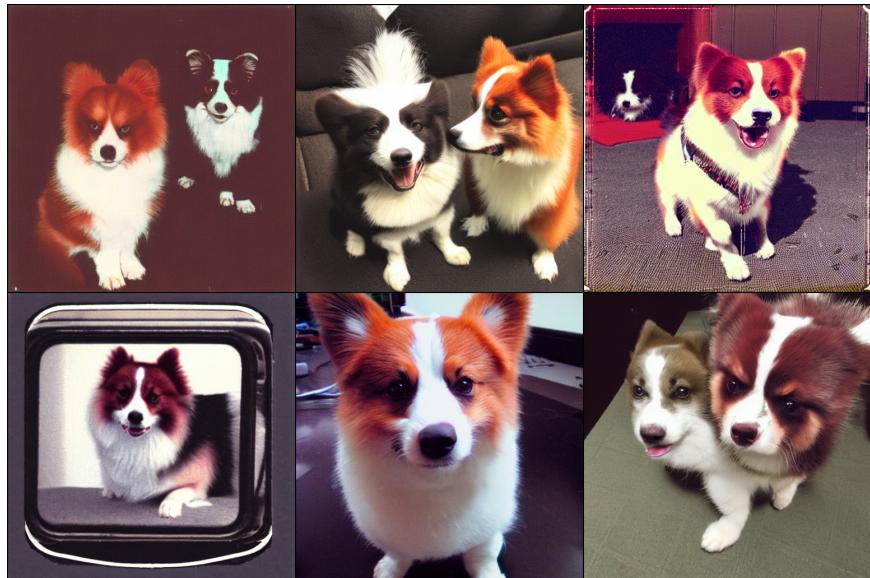
- 5 failed cases:

Image	7_480	7_481	7_483	7_484	7_487
Wrong label	willow_tree	willow_tree	willow_tree	kangaroo	willow_tree

2. What will happen if you simply generate an image containing multiple concepts (e.g., a <new1> next to a <new2>)? You can use your own objects or the provided cat images in the dataset. Share your findings and survey a related paper that works on multiple concepts personalization, and share their method.
- 我訓練 special token <new1> 為 dog 圖片、<new3> 為 cat 圖片：
    - 當我使用 prompt: "A photo of <new3> next to <new1>" 時，我發現生成出的圖片會比較接近 <new3> 的特徵，如下圖：



- 但我使用 prompt: "A photo of <new1> next to <new3>" 時，我發現生成出的圖片會比較接近 <new1> 的特徵，如下圖：



- Multiple concepts personalization paper: Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, Wenhan Luo. (2024). OMG: Occlusion-friendly Personalized Multi-concept Generation in Diffusion Models. ECCV 2024 論文的實作方法如下:

  - Visual Comprehension Information Preparation : 要通過對 cross-attention layers 進行控制，從而保留每個概念的空間佈局和外觀，在 cross-attention layers 中，通過修改像素與文本之間的交互，可以保留圖像中原有的結構和內容，同時添加新的概念。
  - Visual Comprehension Information Preparation : 使用 Concept Noise Blending 方法，以無需進行模型合併的方式整合多個概念，每個單一概念模型負

責生成其特定的概念，在生成過程中，根據不同的概念遮罩將每個概念注入到圖像的不同區域中。

3. Occlusion Layout Preservation：在去噪過程中保留 cross-attention 映射，確保多概念之間的遮擋情況能夠得到有效處理，這樣，即使在不同時間步中生成的噪聲存在變化，最終生成的圖像仍能保持一致的佈局和概念完整性。
- 

## Reference

[DDPM Github repo](#)

[DDIM Github repo](#)

[UNet Conditional Kaggle](#)

[How to add new tokens to an existing Huggingface AutoTokenizer?](#)

[How to add new special token to the tokenizer?](#)

[How to train the embedding of special token?](#)

[huggingface train only new tokens embedding](#)

[How to train new token embedding to add to a pretrain model?](#)

[What is Textual Inversion - Textual Inversion Github](#)

[\[Github rinongal/ textual\\_inversion\] An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion](#)