DLCV hw2 Report

姓名:謝銘倫 學號:**B10502166** 系級:電機四

p1:

1.

I implement classifier-free guidance with the DDPM class. With writing the q_sample formula to adding noise forwardly and p_sample to denoise while sampling, I map the label of MNISTM class to **0-9** and SVHN class to **10-19** such that DDPM model can identify different labels and non-condition cases in classifier-free guidance at once.

The most difficult part is to find the proper w (guidance strength) to make my image conditioned on the labels while sampling. I found that **bigger** w will lead to a better generation, e.g. w = 40 with respect to my device.

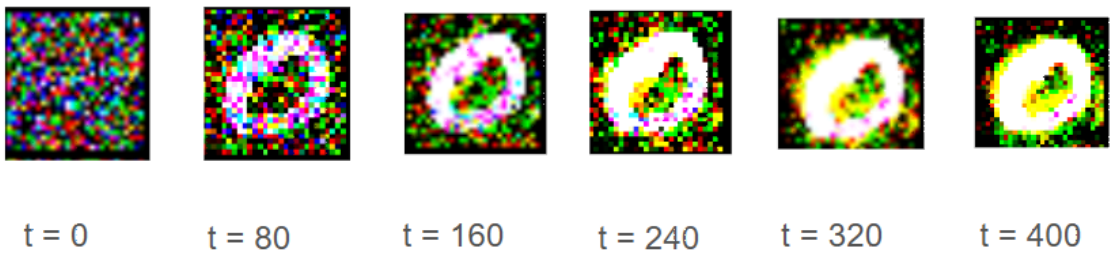
2.

(Fig1: MNISTM dataset sampling 10*10 numbers)

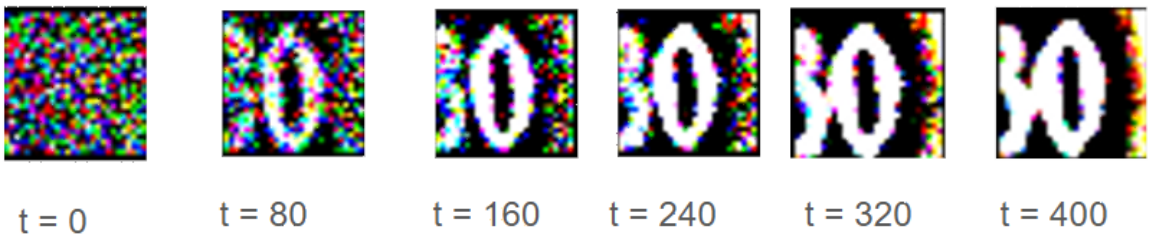
(Fig.2: SVHN dataset sampling 10*10 numbers)

3.

Total timestep: 400



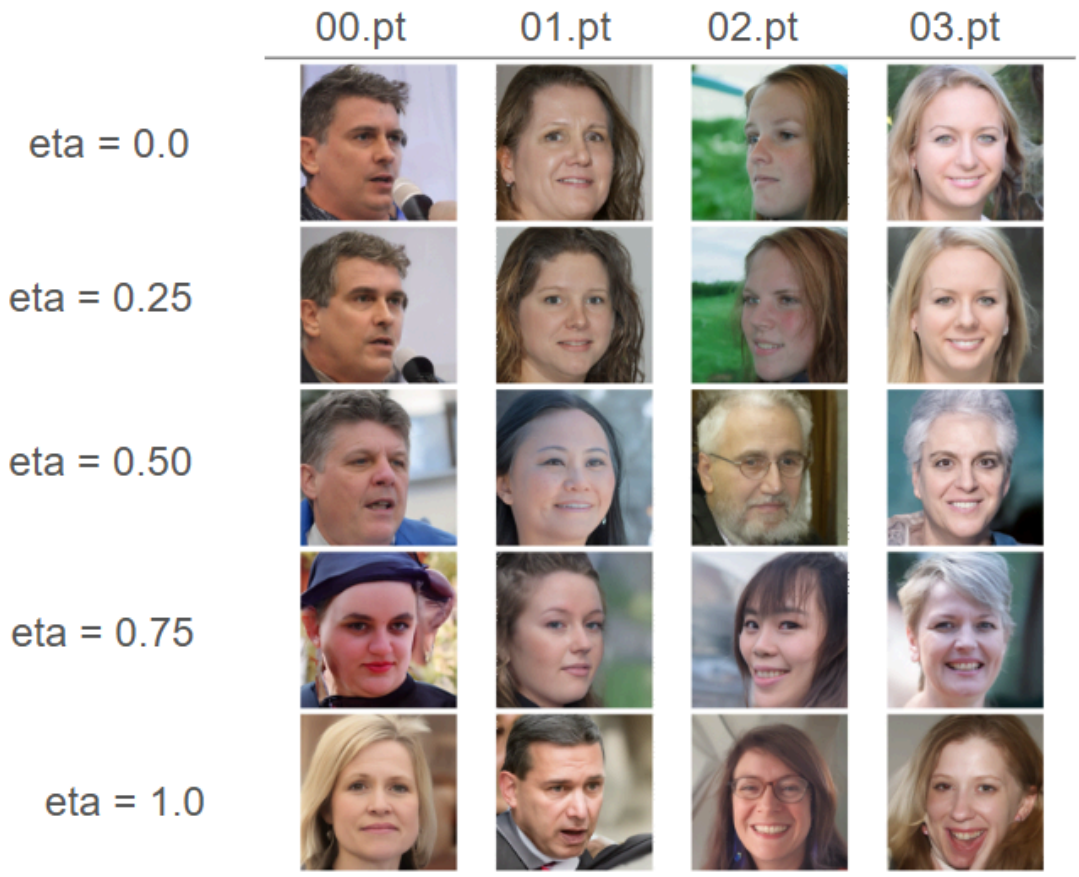
(Fig.3: The number 0 of the MNISTM dataset for each sampling timesteps)



(Fig.4: The number 0 of the SVHN dataset for each sampling timesteps)

p2:

1.



(Fig.5: The DDIM sampling over the eta = [0.0, 0.25, 0.50, 0.75, 1.0] over **00.pt~03.pt**)

2.

(Fig.6: SLERP, the face images of the interpolation of noise **00.pt ~ 01.pt.**)

(Fig.7: Linear Interpolation, the face images of the interpolation of noise **00.pt ~ 01.pt.**)

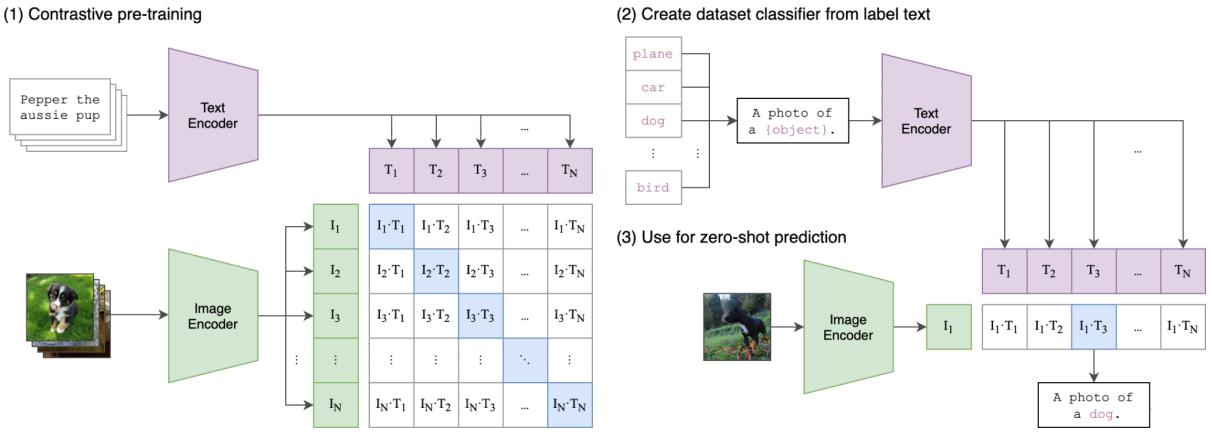
Observation: The linear interpolation will cause the destruction of 00.pt, and simply construct the contour of face of the 01.pt, which looks supernatural.

p3:

1.

The Clip Encoder use both text encoder and image encoder to get each feature and calculate cosine similarity, which reveals how the text and image related to each other.

After obtaining the cosine similarity, we can do classification by finding the highest similarity score to categorize the label of image based on text. That's how zero-shot classification was done after running CLIP Encoder.



(Fig.8: The Structure of CLIP Encoder)

Accuracy: 58.60%

```
(ldm) (base) chrishsieh@LAPTOP-E1117TUI:~/dev/DLCV/dlcv-fall-2024-hw2-chrisPixelCraft/p3_clip$ python3
 clip_zeroshot.py
Using device: cuda
Loaded 50 labels
Evaluating accuracy: 100%|| 2500/2500 [01:28<00:00, 28.26it/s]

Accuracy: 58.60%
```

(Fig.9: The accuracy output of the CLIP zero-shot classification, i.e. 58.6%)

```
Successes:
0_450.png - True/Pred: bicycle @ confidence: 0.8%
0_451.png - True/Pred: bicycle @ confidence: 0.9%
0_452.png - True/Pred: bicycle @ confidence: 0.9%
0_453.png - True/Pred: bicycle @ confidence: 0.9%
0_454.png - True/Pred: bicycle @ confidence: 0.4%
```

(Fig.10: Five success classifications)

```
Failures:
0_467.png - True: bicycle, Pred: willow_tree @ confidence: 0.3%
0_474.png - True: bicycle, Pred: willow_tree @ confidence: 0.2%
0_479.png - True: bicycle, Pred: willow_tree @ confidence: 0.1%
0_486.png - True: bicycle, Pred: pine_tree @ confidence: 0.2%
0_494.png - True: bicycle, Pred: oak_tree @ confidence: 0.6%
```

(Fig.10: Five failed classifications)

2.

By using <new1> and <new2> with the multi-concept personalization: The result is not ideal for representing two concept, we need further methods.

(Fig.13 generated image: A photo of <new1> <new2>)

Paper survey with multi-concept personalization:
The paper *"FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition"* introduces a novel method for generating target images without requiring fine-tuning, contrasting traditional approaches like textual inversion and DreamBooth. FreeCustom is efficient, enabling multi-concept feature learning with significantly reduced training time. This approach employs a U-Net denoising model, where training images first undergo processing through a Variational Autoencoder (VAE) before the forward process. The U-Net then

denoises these images, using Q, K, and V from self-attention layers fed into the main model. Additionally, a mask generator provides conditional input, and through Multi-Resolution Self-Attention (MRSA), the model computes these conditions to produce a latent space image z′z'z′, which is decoded to the final target image.