Debiasing Facial Generation and Reconstruction Tool

去偏見化之臉部生成和臉部重建工具

指導教授：蔣榮先
專題成員：黃霖均
開發工具：PyTorch, Stable-Diffusion-v1-5, Deepface, torch-fidelity, clean-fid

一、　簡介

As generative AI like diffusion models gain widespread use, ensuring fairness and mitigating harmful biases has become paramount. These powerful models can generate highly realistic synthetic data. However, if these synthetic data exhibit societal biases, it may risk perpetuating harmful stereotypes, exacerbating social divides.

As shown in Fig. 1, this project aims to mitigate bias in generative diffusion models by introducing a sensitive attribute learning indicator during training. For instance, the model takes a face image or a natural image as an input, while the model learns how to generate a human face only, and the indicator learns how to distinguish between a face and the others.

The indicator is characterized by non-binary attribute learning and lightweight structure. The indicator utilizes non-binary attribute learning to mitigate biases like gender by considering visual features without being constrained to binary notions, while its lightweight convolutional structure minimally impacts training time.

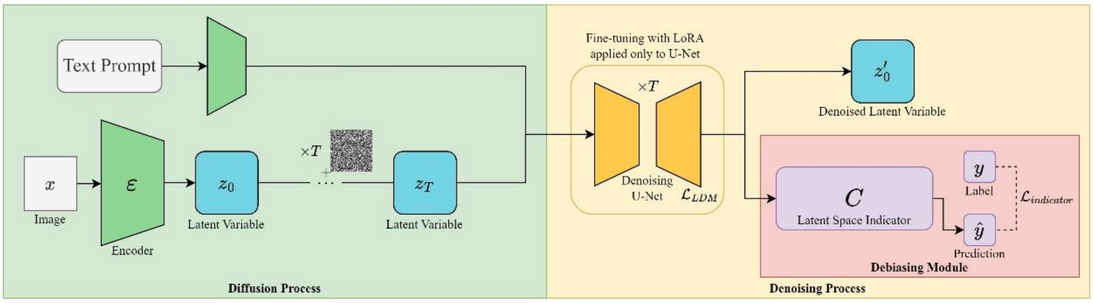

Fig. 1 Debiasing Diffusion Model. The indicator C outputs a value which serves as a key factor in the loss computation, thereby guiding the model’　s learning process.

This project is demonstrated in the Debiasing Facial Generation and Reconstruction Tool. Both a text prompt or a face contour can be input as a guidance to the model, and the model generates a corresponding face image.

二、　測試結果

In our experiments, as shown in Table 1 and Fig. 2, we train models with different debiasing intensity and different datasets, and generate 800 images to count the demographic composition with Deepface in these models.

Training Set Gender Ratio (Male : Female)	Debiasing Intensity	FD(↓)	FID(↓)	IS(↑)
0.25	0	0.14	0.017	3.125±0.242
	0.01	0.11	0.023	3.113±0.177
	0.05	0.08	0.025	3.899±0.263
	0.1	0.1	0.043	3.499±0.234
1	0	0.08	0.015	3.153±0.231
	0.01	0.02	0.015	3.590±0.273
	0.05	0.05	0.016	3.508±0.341
	0.1	0.03	0.026	3.803±0.295
4	0	0.30	0.014	4.029±0.505
	0.01	0.40	0.018	3.880±0.211
	0.05	0.14	0.014	3.971±0.302
	0.1	0.40	0.021	3.518±0.198

Table 1. We test Fairness Discrepancy (FD), Fréchet Inception Distance (FID), Inception Score (IS) of our models. A lower FD score indicates a fairer result, a lower FID score indicates a higher-quality image, and a higher IS indicates better quality and diversity in the generated samples. The best results are highlighted in bold on the table.

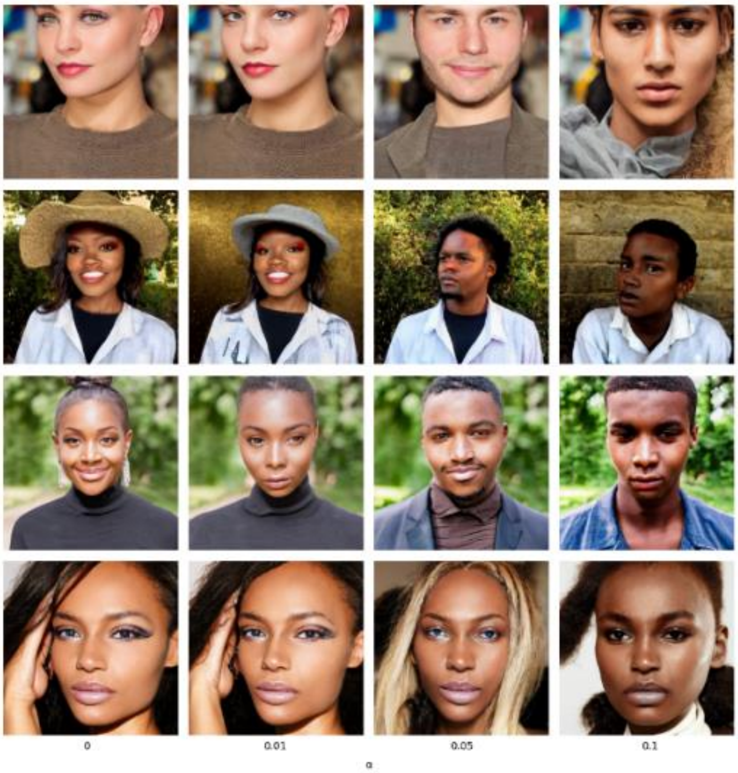

Fig. 2 The figure demonstrates the effects of different α values (intensity) on the debiasing process, using a model fine-tuned with a dataset comprising 20 male and 80 female images. Each row represents images generated with a consistent seed, showing the debiasing effects as α increases from left to right.