DNA-GAN: Learning Disentangled Representations from Multi-Attribute Images

Taihong Xiao, Jiapeng Hong & Jinwen Ma

Department of Information Sciences, School of Mathematical Sciences and LMAM, Peking University

xiaotaihong@pku.edu.cn, jphong@pku.edu.cn, jwma@math.pku.edu.cn

Introduction

Disentangling factors of variation has become a very challenging problem on representation learning. However, existing algorithms suffer from many limitations, such as unpredictable disentangling factors, unable to identify different styles of a certain attribute, lack of identity information, etc. Motivated by DNA double helix structure, in which different kinds of traits are encoded in different DNA pieces, we make a similar assumption that different visual attributes in an image are controlled by different pieces of encodings in its latent representations.

Model

As shown in Figure 1, DNA-GAN is mainly composed of three parts: an encoder (Enc), a decoder (Dec) and a discriminator (D).

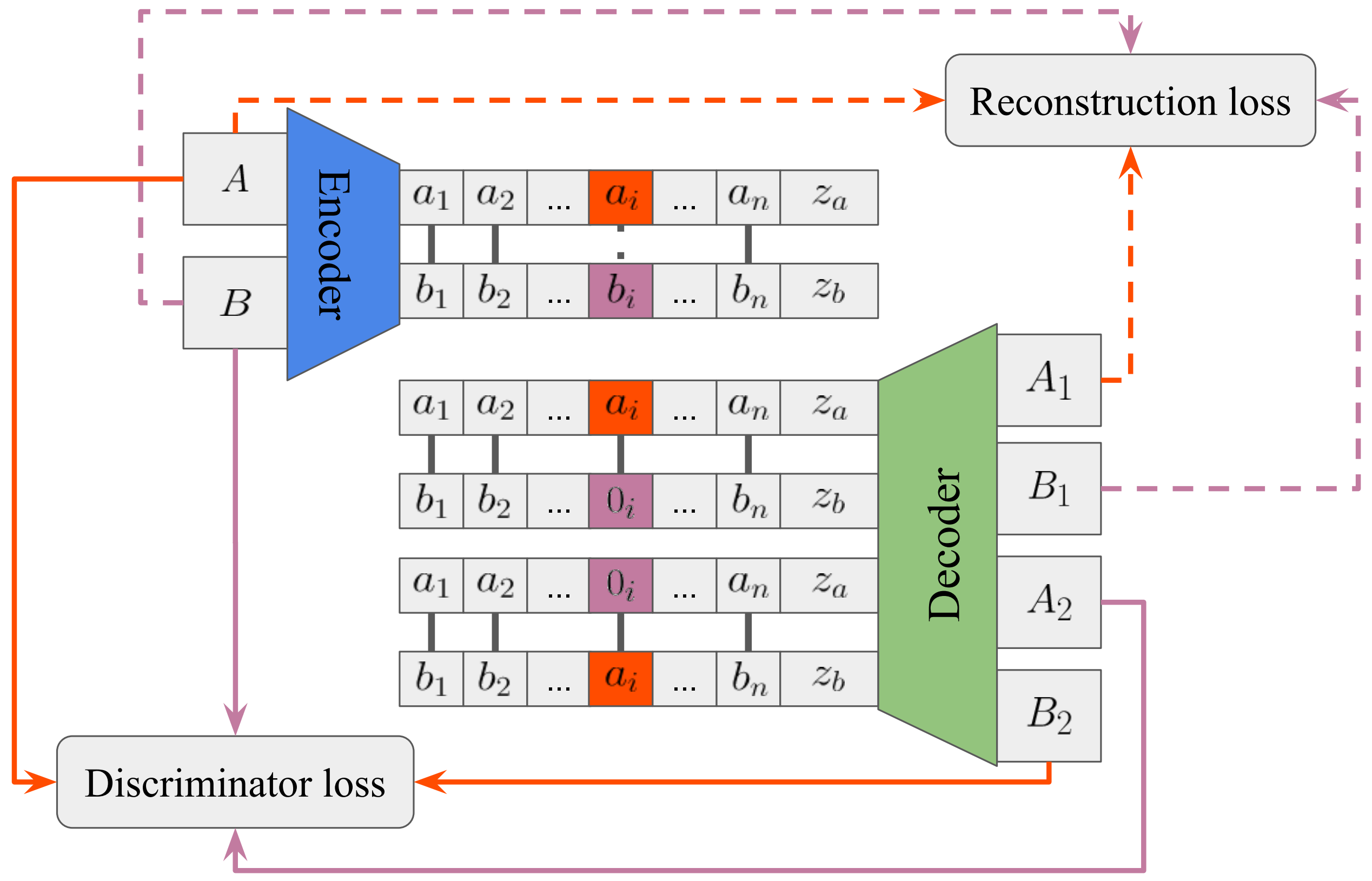
Reconstruction loss
A
B
Encoder
a_1 a_2 ... a_i ... a_n z_a
b_1 b_2 ... b_i ... b_n z_b
a_1 a_2 ... a_i ... a_n z_a
b_1 b_2 ... 0_i ... b_n z_b
a_1 a_2 ... 0_i ... a_n z_a
b_1 b_2 ... a_i ... b_n z_b
Decoder
A_1
B_1
A_2
B_2
Discriminator loss

Figure 1: DNA-GAN architecture.

We focus on one attribute in each iteration. Let's say we are at the i-th attribute. A and B are required to have such labels, $Y^A = (\mathbf{y}_1^A, \dots, 1_i^A, \dots, \mathbf{y}_n^A)$ and $Y^B = (\mathbf{y}_1^B, \dots, 0_i^B, \dots, \mathbf{y}_n^B)$, respectively. The encoder maps the real-world images A and B into two latent disentangled representations.

$$\mathrm{Enc}(A) = [a_1, \dots, a_i, \dots, a_n, z_a], \quad \mathrm{Enc}(B) = [b_1, \dots, b_i, \dots, b_n, z_b]$$

We copy $\mathrm{Enc}(A)$ directly as the latent representation of A_1, and annihilate b_i in the copy of $\mathrm{Enc}(B)$ as the latent representation of B_1. The annihilating operation means replacing all elements with zeros.

By swapping a_i and 0_i, we obtain two new latent representations $[a_1, \dots, 0_i, \dots, a_n, z_a]$ and $[b_1, \dots, a_i, \dots, b_n, z_b]$ that are supposed to be decoded into A_2 and B_2, respectively. Via a decoder Dec, we can get four newly generated images A_1, B_1, A_2 and B_2.

$$\mathrm{Dec}([a_1, \dots, a_i, \dots, a_n, z_a]) = A_1, \quad \mathrm{Dec}([b_1, \dots, 0_i, \dots, b_n, z_b]) = B_1$$
$$\mathrm{Dec}([a_1, \dots, 0_i, \dots, a_n, z_a]) = A_2, \quad \mathrm{Dec}([b_1, \dots, a_i, \dots, b_n, z_b]) = B_2$$

Loss Functions

The encoder and decoder receive two types of losses: (1) the reconstruction loss,

$$L_{reconstruct} = \|A - A_1\|_1 + \|B - B_1\|_1$$

(2) the standard GAN loss,

$$L_{GAN} = -\mathbb{E}[\log(\mathrm{D}(A_2|\mathbf{y}_i^A = 1))] - \mathbb{E}[\log(\mathrm{D}(B_2|\mathbf{y}_i^B = 0))]$$

Omitting the coefficient, the loss function for the encoder and decoder is

$$L_G = L_{reconstruct} + L_{GAN}.$$

The discriminator D receives the standard GAN discriminator loss

$$L_{D_1} = -\mathbb{E}[\log(\mathrm{D}(A|\mathbf{y}_i^A = 1))] - \mathbb{E}[\log(1 - \mathrm{D}(B_2|\mathbf{y}_i^A = 1))]$$
$$L_{D_0} = -\mathbb{E}[\log(\mathrm{D}(B|\mathbf{y}_i^B = 0))] - \mathbb{E}[\log(1 - \mathrm{D}(A_2|\mathbf{y}_i^B = 0))]$$
$$L_D = L_{D_1} + L_{D_0}$$

Experiments

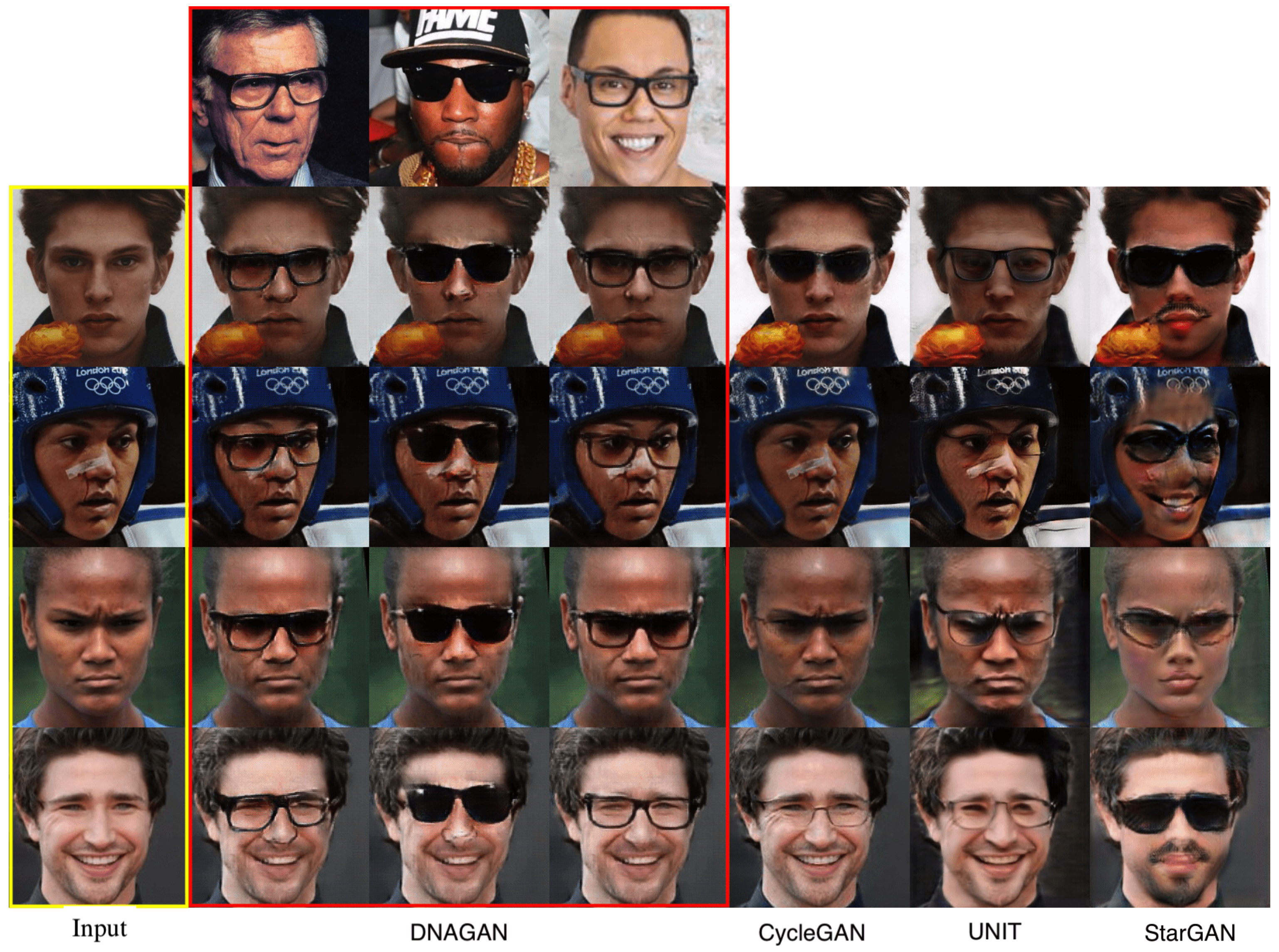

Figure 2: Face image generation by adding different types of eyeglasses. The yellow box indicates the input images, the three on the top are reference images. Our model can add exact the same eyeglasses in the reference images to the input image.

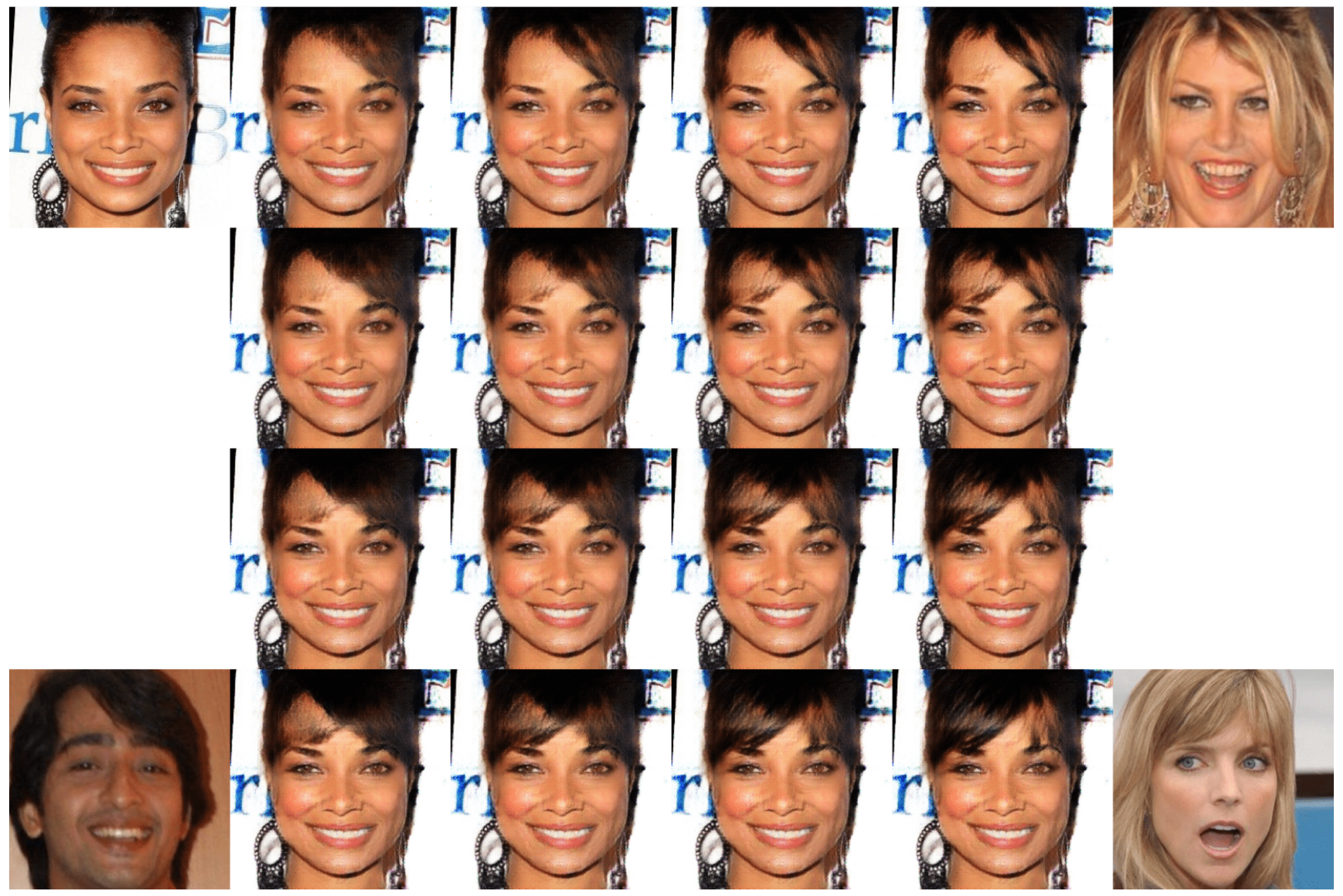

Figure 3: Interpolation of different types of bangs. The top-left image is the input and images at the other three corners are three reference images of different styles of bangs.

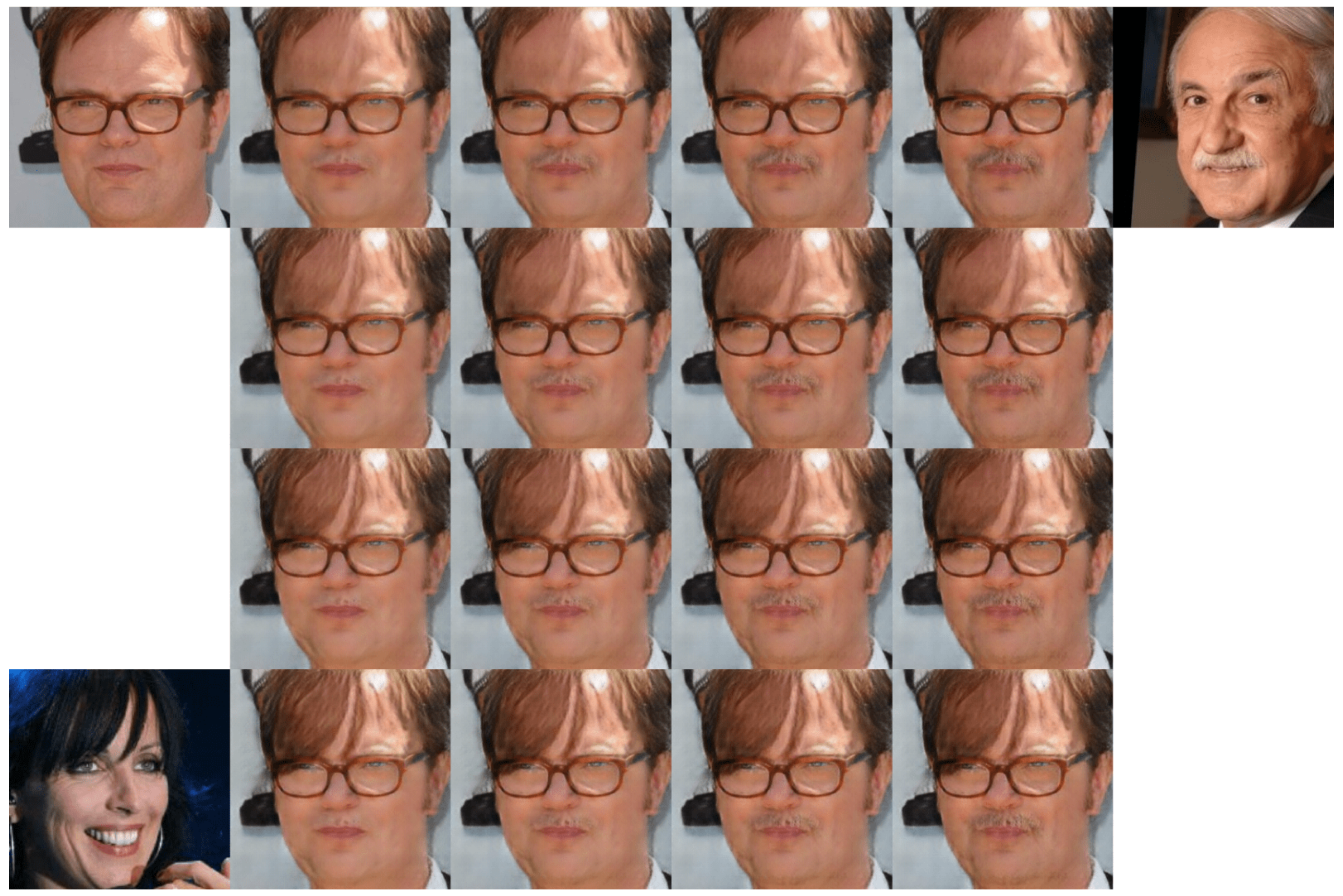

Figure 4: Interpolation of bangs and mustache. The top-left image is the input. The bottom-left is the reference image of the `bangs` attribute and top-right image is the reference image of the `mustache` attribute.

Scan for source code.

https://github.com/Prinsphield/DNA-GAN