AttGAN: Facial Attribute Editing by Only Changing What You Want

Antoine BOHIN, Logan RENAUD

Deep Learning Course
3A-MDS-DL
7th Jan. 2025

Context

Facial attribute editing aims **to edit attributes of interest** (e.g. hair color, accessories, mouth) on a face image while preserving the identity of the person. It has applications in various industries, like entertainment and media generation (e.g. deepfakes, face manipulation in films or photo shooting), but also plastic surgery or security.

With the lack of suitable labeled datasets, current approaches rely on generative models like Generative Adversarial Networks (GANs) and encoder-decoder architectures aiming at decoding latent representation of a given face conditioned on the desired attributes. Some of them attempt to build attribute-independent latent representations. However, these techniques often suffer from loss of facial details, unintended attribute changes due to correlations in the dataset, and the need for multiple models to handle different attribute edits.

Attribute GAN (AttGAN) [1] introduces an improved framework, leveraging an encoder-decoder structure combined with attribute classification constraints, reconstruction learning and adversarial learning at training. It results in more realistic facial transformations, better retention of details and a more flexible model that can handle multiple attributes simultaneously with a single implementation.

Fig. 1: Facial attribute editing with AttGAN

Methodology

AttGAN consists of an **Encoder-Decoder structure** to generate the images, through $x^{\hat{b}} = G_{\text{dec}}(G_{\text{enc}}(x^a), b)$ where we denote a and b the attributes of the original and desired image, and x^a and $x^{\hat{b}}$ the original and final edited image (see below)

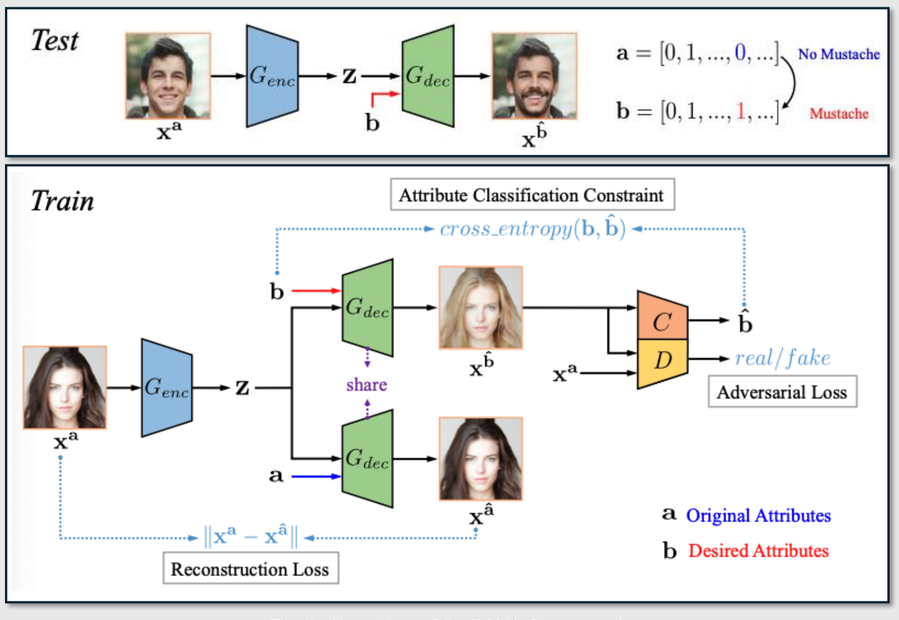


Fig. 2: Overview of AttGAN's framework

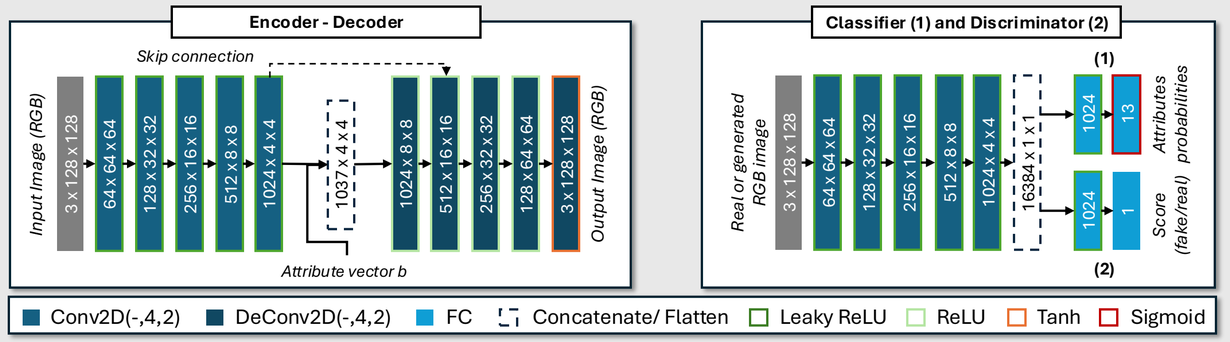
The model is combined with **additional components used during training** to learn how to generate realistic and precise edits:

1. A pre-trained **Attribute Classifier** is used to check whether **the attributes in the output image match the desired ones**. The constraint is based on the binary cross-entropy loss : $L_{cls}^g = E_{x^a \sim p_{data}, b \sim p_{attr}} \left[\sum_{i=1}^{n} -b_i \log C_i\left(x^{\hat{b}}\right) - (1-b_i) \log\left(1 - C_i\left(x^{\hat{b}}\right)\right) \right]$
2. The **Reconstruction Learning** aims at **preserving the attribute-excluding details**. If an image is passed without modification, it should be perfectly reconstructed, hence we use the reconstruction loss: $L_{rec} = E_{x^a \sim p_{data}}\left[|x^a - x^{\hat{a}}|_1\right]$
3. An **Adversarial Learning** framework is employed for **visually realistic generation**. A **discriminator** ensures that the generated images look realistic and indistinguishable from real pictures. An adversarial loss is computed: $L_{adv}^d = -E_{x \sim p_{data}}[D(x)] + E_{x^a \sim p_{data}, b \sim p_{attr}}\left[D\left(x^{\hat{b}}\right)\right]$ for the discriminator and $L_{adv}^g = -E_{x^a \sim p_{data}, b \sim p_{attr}}\left[D\left(x^{\hat{b}}\right)\right]$ for the generator

The overall objective is obtained by combining those different losses.

Chosen Experimentation Settings

We used the CelebA Dataset [2] containing more than 200k celebrity images, with binary annotations for +40 attributes. For practical reasons, we chose to firstly focus only on 13 key attributes. The chosen architectures for the model components are described below:



Attribute injection is performed by concatenating the attribute vector with the encoded feature map before decoding. The model also presents shortcut connections, inspired by U-Net to help preserve fine details during image reconstruction.
We aim to compare AttGAN with models that does not include attribute classification constraint or reconstruction learning, like **VAE [3]** and **IcGAN [4]**. We perform **qualitative evaluation** to analyze **identity preservation, image quality** and **attribute correctness**.

Results

Our implementation of AttGAN **successfully modifies facial attributes while preserving identity and image consistency**. The model accurately applies attribute edits (e.g. changing hair color, adding eyeglasses…), without introducing artifacts.
Even with a lower number of training epochs (45) compared to the paper (200), the model produces **realistic and high-quality edits**, confirming the robustness of its encoder-decoder architecture and adversarial training approach.

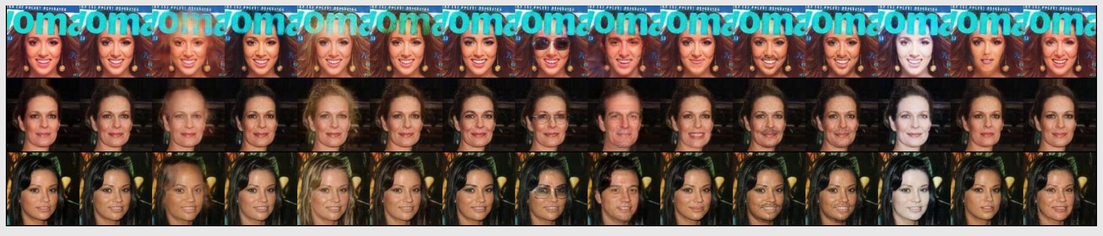
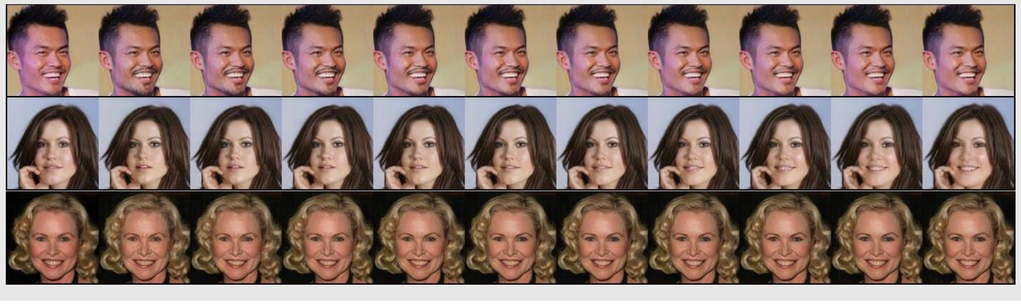


Fig. 4: Single attribute editing on CelebA test images. From left to right: *Input, Reconstruction, Bald, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Male, Mouth Slightly Open, Mustache, Beard, Pale Skin, Smiling, Young*

We also explored the model's **capability to generalize beyond binary attribute values** and use **continuous intensity levels for attribute editing**. While AttGAN is originally trained with discrete 0/1 labels, we observed that it **can naturally handle gradual attribute modifications** during testing, **without requiring any architectural changes**.

Fig. 5: Intensity control for the attributes *Beard, Smiling, Young*



References

[1] He et Al. ; *AttGAN: Facial Attribute Editing by Only Changing What You Want* ; https://arxiv.org/abs/1711.10678
[2] CelebFaces Attributes Dataset (CelebA) ; http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3] Larsen et Al. ; *Autoencoding beyond pixels using a learned similarity metric* ; https://arxiv.org/pdf/1512.09300
[4] Perarnau et Al. ; *Invertible Conditional GANs for image editing* ; https://arxiv.org/pdf/1611.06355