**Abstract**

*The recent advancement in text-to-image generation is mainly due to multi-modal representation learning. While many neural-network-based models for the task inherit the data-hungry aspect of deep learning demanding a large amount of human-annotated image-caption pairs, contrastive learning allows representation learning in a self-supervised manner. Here we study hard and present a novel method which outperforms everything!*

# 1. Introduction

An Image may contain a vast number and variety of attributes. An agent with the capability of processing visual information can categorize attributes, classify images with respect to the category, and even can imagine a scenery having a certain collection of visual attributes.

Image attribute manipulation is the task of translating an input image to a new realistic image that has a certain subset of attributes modified as desired, while the rest remains as the original. In general, it is a multi-domain task, which requires learning the translation function for numerous attributes.

One of the challenges of the image attribute manipulation task rises from the fact that different parts of the image can contribute to different attributes. An attribute can be local, global, or even an abstract characteristic of the image Considering facial images, for example, some attributes including hair colour and eyeglasses may be local, while others such as the gender or the ethnicity of an individual may be a result of the composition of local attributes. The dependency amongst attributes makes the disentanglement of attributes to be difficult.

Another challenge is that the quality of synthesized visual content is heavily reliant on the relations among its components, so an attempt of manipulating a single attribute may be generating unrealistic images. For example, baldness or moustache may be a local facial attribute, especially for a classification task, while it exhibits different patterns in each gender, and is more expressive among males. Such feature dependency essentially calls to learn different translations on the embedding space for each attribute combination which potentially has exponentially many possibilities.

Despite the difficulties of the task, there has been great progress from the generative adversarial network (GAN) based models. In particular, the decomposition-based strategies that control the latent embeddings of attributes to modify image attributes have been proposed [[4](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.87elmpkm0kgw), [5](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.km7l95sfzr2l), [6](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.s915ijxsit0s)] showing that the strategy can result in attribute modification with high controllability of the generated image.

In this project, we extend the previous work by studying the attribute representation learning method and its impact on the image attribute modification task. More specifically, we aim to study how the different contrastive learning losses and strategies can improve the interpretability and controllability of the GAN-based model.

# 2. Related Works

## 2.1. Generative Adversarial Network

### GAN in attribute manipulation task

~~The survey on GAN-based facial attribute manipulation can be found in [~~[~~1~~](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.9y2u5phwygr0)~~].~~ Most GAN-based attribute manipulation models take the source image and target attributes as input. The generator of the model often contains an encoder, which obtains the embedding of the input image. The target attribute can be either directly fed into the decoder with the source image embedding (as in AttGAN [[2](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.x46o1qfd6bnr)] or StarGAN [[3](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.b1e6isk7tvbk)]) or used to create the embedding of the target image to be generated.

Taking the target image representation as a vector in the latent embedding space sampled with the condition of desired attributes seems to promise better interpretability and finer controllability. DNA-GAN [[4](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.87elmpkm0kgw)] attempted to learn the image representation as a concatenation of segments containing information on each attribute. Recently, GAN-Control [[5](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.km7l95sfzr2l)] demonstrated the efficacy of the contrastive learning approach in the disentanglement of attributes with improved performance and controllability. Further extension of ideas in this direction allowed viewing the attribute manipulation as a problem of finding the transformation function on the latent space conditioned to desired attribute changes. The idea of navigating latent space for attribute manipulation can be found for example LatentCLR [[6](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.s915ijxsit0s)]. Such an idea is particularly useful on generating high-resolution images since it directly uses the generator of pre-trained large-scale GAN models.

In this project, we investigate the effect of different contrastive learning strategies on the training of AttGAN [[2](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.x46o1qfd6bnr)] and StarGAN [[3](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.b1e6isk7tvbk)].

### AttGAN

To Do

### StarGAN

To Do

## 2.2. Self-supervised Representation Learning in Vision and Image Processing.

Contrastive learning is a general framework of learning representation space of the given data. The main idea of contrastive learning is that the metric or the distance measure must correspond to the similarity of data points. Alternative to the metric learning view, it is possible to connect the framework to the idea of mutual information maximization. By maximizing the estimated mutual information (MI) of different views of the input data, unsupervised or self-supervised representation learning ensures more generalizability can be achieved. The difficulty of estimating MI can be mitigated by using the Hilbert-Schmidt Independence Criterion (HSIC). In the vision community, the effectiveness of the contrastive learning framework on self-supervised visual representation learning has been empirically demonstrated, namely by simCLR, Barlow Twins, and BYOL.

# 3. Main 1: Descriptions of Proposed Models

# 4. Main 2: Strategies for Contrastive Learning

# 5. Experiments

## 5.1. Datasets and Metrics

For the training and evaluation of the model, CelebFace Attributes (CelebA) [[11](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.pqv8q8bxwnx)] is used. CelebA is probably the most widely used large-scale face dataset in attribute manipulation studies. The dataset contains local, global, and abstract attributes of 40 kinds annotated on 202,599 face images and 10,177 identities. The dataset can be downloaded at the site[[11](https://docs.google.com/document/d/1-RXU84Ixp2Avdjn4cqfY2wjjMPmVEhfsaRBLLfQc7_M/edit#bookmark=id.pqv8q8bxwnx)] using Google Drive. We preprocessed the data using the Torchvision module to resize the image from 178x218 to 128x128 and normalize the data to make it zero mean and fixed standard deviation.

## 5.2. Implementation Details

Code references(e.g. GitHub links)

AttGAN: https://github.com/elvisyjlin/AttGAN-PyTorch

StarGAN:

<https://github.com/yunjey/stargan>

### AttGAN

Since the pretrained model of AttGAN is on a different set of attributes, we trained AttGAN from scratch to obtain a baseline result. We adapt the majority of settings for training AttGAN to get the best results after 10 epochs (~182000 iterations).

Following the same principle from StarGAN below, we refactor the original implementation of the Genrator to additionally retrieve the embedding of the input image. We also adapt the training cycles to read only the number of our desired attributes. The specific implementation will be described in the StarGAN section.

We also implemented AttGAN with Mutual Information Maximization (MIM) contrastive loss. However, with the limited time settings, the modification is still underdeveloped.

### StarGAN

Our study involved training StarGAN for 200000 iterations on five attributes: black hair, blond hair, straight hair, wavy hair, and male. To optimize the generator and the discriminator, we set the learning rates to 0.0001 and utilized beta values of 0.5 and 0.99 for the Adam optimizer.

The base structure of StarGAN does not have a separate encoder and decoder that we can retrieve from the generator. To address this limitation and facilitate our approach, we modified the base code of StarGAN to incorporate a separated encoder and decoder component. This modification enabled us to obtain the embedding of the input image.

During the training cycle, the discriminator of StarGAN takes images as inputs and outputs the attributes classification of its images. The generator takes an image, could be real or fake, and an attribute labels as input and returns an image.

In our implementation of the infoNCE loss onto the training cycle of StarGAN, the attribute representation is a critical component. With our modified structure, this can be computed by incorporating both the identity embedding and the embedding of the input image with a target attribute.

The idea of how we can compute the infoNCE loss is as follows. We start off by selecting two batches of images of different individuals with the same attributes during each training cycle iteration. Then we compute the attribute representation of these two batches and use this to form the similarity matrix. The softmax function is applied to the similarity matrix, and the mean of the diagonal entries is taken to obtain the infoNCE loss. Finally, the infoNCE loss is added to the base total loss of the generator and used to update the generator parameters.

### StyleGAN2

## 5.3. Results

### Attribute Manipulation Results

1. Before and after infoNCE
2. Generate images from the identity embeddings

Original | Reconst.(AttGAN) | Reconst(StarGAN) | Reconst. (StyleGAN2)

Original | Attr1.(...GAN w/o InfoNCE) | w/ InfoNCE

### Infonce StarGan



**Normal StarGan**

AttGAN:  
Original: image 1 is the reconstructed image, the rest follows the patterns of the

<https://drive.google.com/drive/folders/1L3k6Hh7syJpvAAYli9kglaaLS3RpFi6K?usp=share_link>

With InfoNCE: <https://drive.google.com/drive/folders/1g3OFwMLymW95wQ_zpkyLhsQo6rCvZ3CF?usp=share_link>

### 

### Classification Metrics

Original StarGAN: Accuracy score 0.544

F1 score 0.807

InfonceLoss StarGAN: Accuracy score 0.515

F1 score 0.782

Original AttGAN: Accuracy score 0.503, F1 score 0.78

InfonceLoss AttGAN: Accuracy score 0.556, F1 score 0.75

## 5.4. Qualitative Analysis

Observation: when dealing with hair color, the original StarGAN sometimes tends to over colorize the images. An example of this is with the blond hair feature with a black background, the surrounding of the person gets filtered out by yellow color. When dealing with bright colored hair, the base StarGAN seems to have problems with modifying it to be black. Both of these issues seem to be lessen after

# 6. Conclusion

# 

# Contributions

W.C. conceived of the project idea. Each team member will choose one of the GAN-based models for encoding and image generation and explore various contrastive learning loss and strategies. The possible choices of the base model and contrastive learning losses are discussed above. All project members will discuss the results and contribute to the final manuscript.

# References