Spring 2023 Boston University

**CS 585: Image and Video Computing**

**Final Project Topic Proposal**

Wonyl Choi U29967430, Duc Minh Nguyen U17531382, and Quan Pham U52532940

# Description of the Final Project Topic

## **Topic title**

## Contrastive Representation Learning For Image Attribute Manipulation

## **Objective**

The main goal of the project is to investigate the effect of the contrastive representation learning method on the performance of image attribute manipulation. More specifically, the main objectives of the projects are as follows:

* Understand and build the generative adversarial network (GAN) based model for the image generation and image-to-image translation task
* Study the effect of different contrastive learning schemes for image attribute manipulation.
* Measure the performance of the model and perform the ablation study for each design choice made.

## **Summary**

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 result the generation of an unrealistic image. 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 have been great progresses from the generative adversarial network (GAN) based models. In particular, semantic representation decomposition based strategy that controls the latent embeddings of attributes to modify image attributes have been proposed ([[4](#87elmpkm0kgw)], [[5](#km7l95sfzr2l)], [[6](#s915ijxsit0s)]), showing that the strategy can result 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 aims to study how the different contrastive learning losses and strategies can improve the interpretability and controllability of the GAN based model.

**Background**

The survey on GAN-based facial attribute manipulation can be found in [[1](#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](#x46o1qfd6bnr)] or StarGAN [[3](#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](#87elmpkm0kgw)] attempted to learn the image representation as a concatenation of segments containing information on each attribute. Recently, GAN-Control [[5](#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](#s915ijxsit0s)]. Such idea is in particular useful on generating high-resolution images, since it can directly use the generator of pre-trained large-scale GAN models.

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 that 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 the self-supervised visual representation learning has been empirically demonstrated, namely by simCLR, Barlow Twins, and BYOL.

# Related Papers

In this project, we investigate the different contrastive learning methods to different base models, which all are variants of GAN. For the image attribute manipulation task, AttGAN [[2](#x46o1qfd6bnr)] and StarGAN [[3](#b1e6isk7tvbk)] could be considered as the good baseline models. Those models fits to the objective of the project, since it does not utlize the contrastive representation learning.

Another model that we can use as the baseline model is the StyleGAN2 [[7](#xh9pgnpdhazv)]. The pre-trained StyleGAN2 has been used for realizations of several semantic representation decomposition based strategies ([[8](#te7mz220zd94)], [[9](#ty7ukpki40v9)], [[10](#6681ey60752a)]).

# Datasets

For the training and evaluation of the model, the following datasets will be used: CelebFace Attributes (CelebA) [[11](#pqv8q8bxwnx)], Flickr-Faces-HQ Dataset (FFHQ) [[12](#35ud44mwdxja)], and Oxford-102-flowers [[13](#35ud44mwdxja)].

CelebA is probably the most widely used large-scale face dataset in the attribute manipulation studies. The dataset contains local, global, and abstract attributes of 40 kinds annotated on 202,599 face images and 10,177 identities. FFHQ dataset contains 70,000 images crawled from Flickr, and can be used for unsupervised pre-training stage.

Oxford-102-flowers is less complex dataset containing about 6,000 images of 102 categories of flowers, which will be used for testing light-weight model for the proof of concept.



Fig. Sample image and attributes from CelebA dataset[[1]](#footnote-0)

| Dataset | Training Set | Validation Set | Tags |
| --- | --- | --- | --- |
| CELEB-A | 202,599 | 162,770 | 40 Attributes (Binary) |
| FFHQ | 60,000 | 10,000 | N/A |
| Oxford-102-flower | 5,171 | 1,822 | 102 Class Labels |

# Contributions

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

# References

[1] Liu, Yunfan, et al. "GAN-based Facial Attribute Manipulation." *arXiv preprint arXiv:2210.12683* (2022).

[2] He, Zhenliang, et al. "AttGAN: Facial attribute editing by only changing what you want." *IEEE transactions on image processing* 28.11 (2019): 5464-5478.

[3] Choi, Yunjey, et al. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[4] Xiao, Taihong, Jiapeng Hong, and Jinwen Ma. "DNA-GAN: Learning disentangled representations from multi-attribute images." *arXiv preprint arXiv:1711.05415* (2017).

[5] Shoshan, Alon, et al. "Gan-control: Explicitly controllable gans." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[6] Yüksel, Oğuz Kaan, et al. "LatentCLR: A contrastive learning approach for unsupervised discovery of interpretable directions." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[7] Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[8] Shi, Yichun, et al. "Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[9] Collins, Edo, et al. "Editing in style: Uncovering the local semantics of gans." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[10] Nitzan, Yotam, et al. "Face identity disentanglement via latent space mapping." *arXiv preprint arXiv:2005.07728* (2020).

[11] Liu, Ziwei, et al. "Deep learning face attributes in the wild." *Proceedings of the IEEE international conference on computer vision*. 2015.

[12] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks. arXiv e-prints." *arXiv preprint arXiv:1812.04948* (2018).

[13] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008.

1. The image is obtained from https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html [↑](#footnote-ref-0)