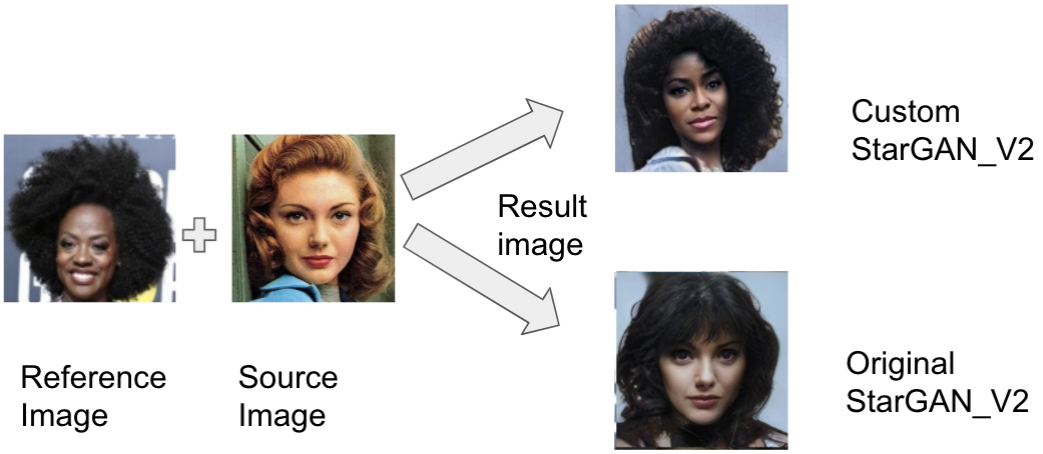
Project: Custom StarGAN
Members: Tabitha Oanda
Advisor: Sarah Bargal

Data

The data for this project had been collected ahead of time. It contains images of faces from 4 different domains: Black Women, Black Men, White Women, White Men
The data used for training the GAN is available online on the SCC and will be made public after the completion of the project. The data for the experiments performed is also on the SCC due to the size limitations on GitHub. The python scripts used for the different experiments are currently available on GitHub in my project repository.

Problem Statement

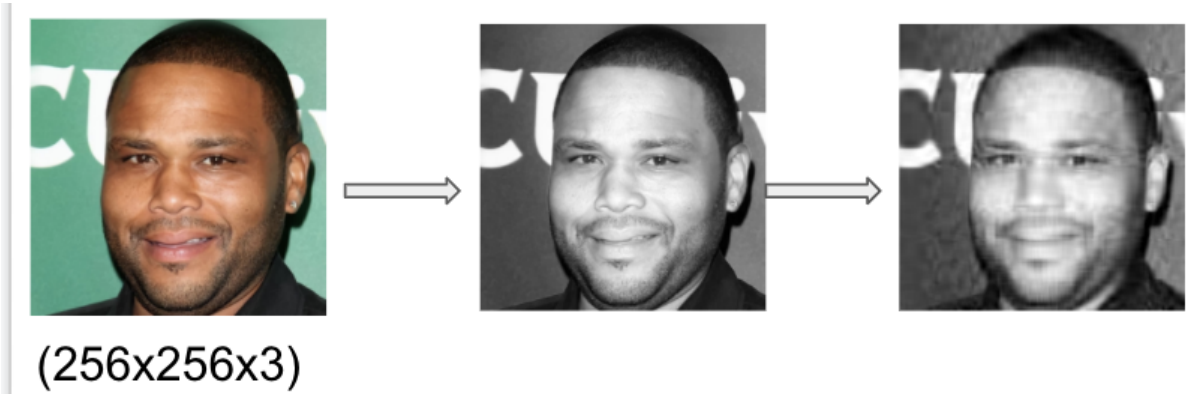
Can I quantitatively measure an improvement or discrepancy between the performance of two GANs.These GANs are custom_StarGAN_V2 and original_StarGAN_V2. In order to measure the performance, I would need to examine the similarity of the generated images to the reference domain. To do this I trained a classifier using KNN and Neural Networks to classify the GAN generated images into their respective domains. The images from the GAN with the smaller classification error indicate better performance of the respective GAN.



Experiments and Results

1. Classifier on KNN

I decided to train a classifier based on the KNN algorithm Ilearned in class. Because KNN fails with high dimension inputs, I decided to downsize my data using SVD. The images went from (256*256*3) vectors to (256*20) vectors. I wanted to keep only 20 singular values because with 20 singular values it is possible to still visualize the image. Below is a projected version of the data on a 256*256 matrix.



After training the classifier based on the rank_k_stargan.ipynb notebook I got accuracy levels below 0.4 despite trying different values of K (2,5,10). I ultimately decided that it was not feasible to measure the performance of the GANs on my KNN model since it had such poor performance on my data.

2. Classifier Using Neural Networks on Colored Images

I decided to train a classifier using a deep neural network. Neural networks have a better way of understanding complex data like images because I don't explicitly define the features to be used. The python script used was classifier.py which is available on GitHub.My final classifier had a validation accuracy of 96% which I found satisfactory.

In order to perform this experiment, I ran my GAN model and the original StarGAN_V2 model on a set of source and reference images belonging to each domain. I cropped these generated images from the matrix and saved them in a folder to be used when the

3. Classifier Using Encoded Images

The plan for this experiment was to use a deep learning model, an autoencoder, to compress the images in order to use them with simpler classifiers i.e. KNN, logistic regression. By the time the fourth deliverable was due. I had successfully trained 4 autoencoders (one for each domain). Here are some of the results from the various autoencoders from the different domains. I am only showing the decoder output because the inputs were identical. The auto encoder was trained using the auto_[domain].py files available on github and the image data available on the SCC.

My advisor and I decided to abort using a classifier trained on encoded images in order to get

more feedback (evaluation metrics) from other models that had been trained on other data besides our own.

4. Results

I decided to perform a gender and skin-tone match test on the images generated from both the original and custom StarGAN_V2 models. The results on the tables below show the average of 4 tests involving 64 unique images generated from each model. The files I use for these tests are Skin_n_gender.ipynb, classifier_predict.py, and the data files are in the test_orig.zip. (All of these are on GitHub)

Gender Accuracy

In this test measure the accuracy of openCVs gender classifier on the generated images. I assume any errors in prediction result from a poorly generated image as opposed to a potentially inaccurate model. This metric was inspired by the Gender Shades paper .

	Custom GAN (gender accuracy)	Original StarGAN_V2 (gender accuracy)
Test 1	0.26	0.25
Test 2	0.47	0.44
Test 3	0.33	0.44
Test 4	0.38	0.56

Skin Tone and Gender Accuracy

I trained a classifier based on real images from CelebA and CelebBlack. The different classes were: Black women, Black men, White Women, White men just as is the case in my domains in StarGAN.

Test	Number of Domains	Custom StarGAN_V2	Original StarGAN_V2
Test 1	4	73%	55%
Test 2	4	64%	45%
Test 3	4	55%	45%
Test 4	2	73%	47%

Skin Tone Accuracy

In this test I create a bounding box around the cheek or forehead and then get the average values of the pixels enclosed in the bounding box. Afterwards I calculate the euclidean distance between the average pixel values of the generated image and the average pixel values of the reference image. In each test I calculate the average euclidean distance between the real reference image and the generated images belonging to that domain. The distance gives an error metric.

(Notation: BW - Black women, BM - Black men, WW - White women, WM - White men)

	Custom StarGAN_V2	Original StarGAN_V2
Test 1	BW : 33.2 BM : 18.3 WW: 23.7 WM: 50.9	BW : 94.7 BM : 204.4 WW: 83.2 WM: 51.8
Test 2	BW : 108.1 BM : 139.6 WW: 120.9 WM: 196.0	BW : 155.0 BM : 196.3 WW: 45.6 WM: 32.5
Test 3	BW : 44.9 BM : 40.0 WW: 28.1 WM: 30.4	BW : 92.3 BM : 124.6 WW: 32.2 WM: 9.9
Test 4	BW : 38.6 BM : 46.9 WW: 59.7 WM: 58.3	BW : 120.2 BM : 145.8 WW: 51.8 WM: 24.7
Overall Average:	BW : 56.2 BM : 61.2 WW: 58.1 WM: 96.4	BW : 115.6 BM : 167.8 WW: 53.2 WM: 29.7

Discussion

I envision this model being used to create avatars that can be used for artistic purposes. Our proposed dataset and data preparation methods can be used to improve visual outcomes for domains with specific visual attributes (eg. gender, race).

Part 1 : Gender Transfer Performance

The original StarGAN outperforms the custom model when it comes to gender adaptation using the Adience classifier. I reasoned that this is because the custom StarGAN takes specific aspects of style from the reference image which might not be highly represented in

the ‘female’ class of the Adience classifier. Another reason for the poorer performance of the custom model is that the generated images do not always look like human faces; the original model is a lot better in terms of face construction.

Part 2 : Skin Tone and Gender Performance
When it comes to a model trained on both race and gender (my custom classifier) the custom model performs better. Both the custom and original model were trained on the same real data that trained the classifier. I can explain the better performance of the custom model as an improvement in skin tone transfer from the Black to the White domain.

Part 3: Skin Tone Performance
In terms of skin tone transfer accuracy my custom model outperforms the original StarGAN model on all Black Domains and the performance on the White women domain is comparable. The original model performs better than the custom model on the White men domain. These results validate the results in part 2 of this discussion. I credit the better performance of the custom model to having more diverse images in the training set and to creating more homogenous image domains. I believe the latter allowed the model to learn more complex features like skin tone and even hair texture (although we don’t test for this) because it had more of these examples in the CelebBlack dataset.

5. Academic Paper

Because this project was a continuation of a project that started in the summer some sections of the paper relate to work I had done previously. The final draft of the paper will have the sections clearly marked out and the results from all the experiments (including the classifier based on GANs). At the moment it contains the parts: Abstract, Introduction, Related Work, Experiments, Discussion. The draft of my paper is linked here

My Advisor and I have submitted the abstract of this paper to the ACM Conference on Fairness, Accountability, and Transparency 2022. In terms of CS 506 the paper is done, however it will need to be reformatted and edited several times to meet the standards and format required for the conference. The CelebBlack dataset will be released along with the paper whether we get published or not.