StyleCLIP: Exploration of Ethnic Bias introduced by a dataset's ethnicity distribution

Darshee Machhar
Rutgers University
dm1639@scarletmail.rutgers.edu

Dhruv Satyapanthi
Rutgers University
ds1990@scarletmail.rutgers.edu

Kunjan Vaghela
Rutgers University
kv353@scarletmail.rutgers.edu

Manan Shukla
Rutgers University
ms3418@scarletmail.rutgers.edu

Mitul Shah
Rutgers University
ms3518@scarletmail.rutgers.edu

Pankti Nanavati
Rutgers University
pn266@scarletmail.rutgers.edu

ABSTRACT

"StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery" is a paper that has derived a way of modifying images through text prompts by leveraging the power of StyleGAN and CLIP. In this paper, the Contrastive Language-Image Pretraining (CLIP) model is used in order to develop a text-based interface for StyleGAN image manipulation. There are three techniques proposed by them viz, Latent Optimizer, Latent Mapper and Global Direction. Amongst these approaches, Latent Mapper turns out to be the most preferred (generates output while maintaining the original characteristics) approach. This paper's authors trained Latent Mapper on the CelebA-HQ dataset which has non-uniform ethnicity distribution. So, we aim to analyze the model's bias towards the ethnicity distribution of the dataset. In order to do so, we trained their model on our newly prepared dataset - with dominant non caucasian facial pictures to get an estimate of ethnicity bias.

KEYWORDS

CLIP, StyleGAN, Latent Space, Disentanglement, Image Manipulation, Prompt Engineering

1 INTRODUCTION

CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on 400 million (image, text) pairs extracted from the web. CLIP consists of two encoders: one for image and one for text. These encoders are trained to map images and corresponding text to a common latent space. The distance between encoded image and text in the CLIP space will be contingent to the similarity score provided by CLIP is actually the inner product of the image and text embeddings within that latent space. StyleGAN is a type of GAN(Generative Adversarial Network). GAN is a state-of-the-art image generator that produces synthetic images with extremely high quality. But StyleGAN is much more than just a great generator. Its latent spaces are disentangled to offer unique editing capabilities. The paper proposes a method that employs latent space to edit images. This method manipulates semantic attributes such as age, eyelashes, brows, hair color, etc. It was originally trained on CelebA-HQ which has dominant caucasian images.

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

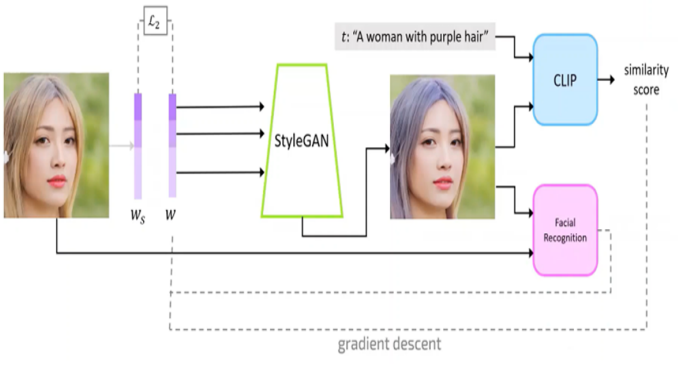
2 LITERATURE REVIEW OF THE PAPER

Along with StyleGAN and CLIP, pretrained models e4e and ArcFace are also used. Encoder4editing (e4e) is used to invert a given image into a style vector. A style vector is basically an embedding of an image within a style latent space using which the StyleGAN can generate an image close to the original image. Hence, we only need to provide the model with images of faces without any text annotation.
Three different techniques are proposed:

2.1 Latent Optimizer

The latent optimization approach in which the latent code is optimized with maintaining the source's identification and moving in the text's direction.

$$\arg\min_{w \in W+} D_{\text{CLIP}}(G(w), t) + \lambda_{\text{L2}} \|w - w_s\|_2 + \lambda_{\text{ID}} \mathcal{L}_{\text{ID}}(w)$$

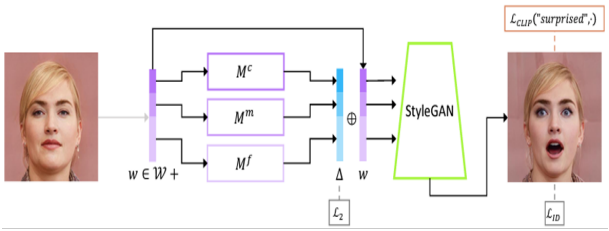


Every approach starts by inverting the image (using e4e) and finding the latent code which corresponds to the input image. For the optimization, it duplicates the latent code and inserts it into the pre-trained StyleGAN to obtain an image. Then, it measures the similarity between this generated image and the input text using CLIP. It uses this similarity score to update the duplicated code

through gradient descent. Additionally, the optimized latent code (w) should stay close to the original latent code (ws) to preserve the input image. For this, the L2 loss is used between the original code and the optimized code. To further preserve the identity of the person in the input image, the identity loss which is measured through a pre-trained face recognition network is used.

2.2 Latent Mapper

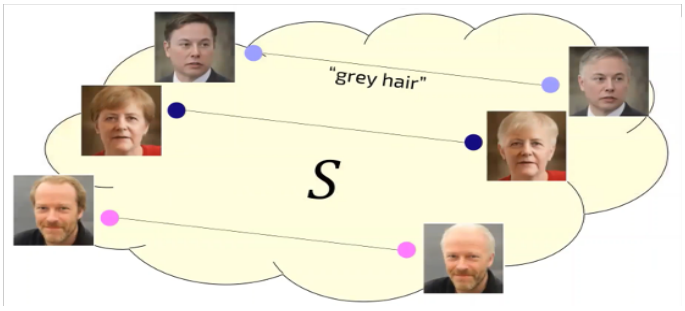
In this approach, a network that receives a latent code is trained to calculate the offset.



To train a latent mapper, again the images are inverted. Their latent code is then inserted into a network and returns an offset. This network is called mapper which is text specific. These offsets are then added to the input latent code and inserted into the pre-trained StyleGAN which generates the output image. This is the forward pass of the mapper.
Here, clip loss is used to measure the cosine distance of the generated image with the text, L2 loss to make sure that the offset is small and identity loss to encourage the mapper to preserve the identity. These losses help to generate images that are of the same person, and only a particular attribute or emotion is changed in the output.

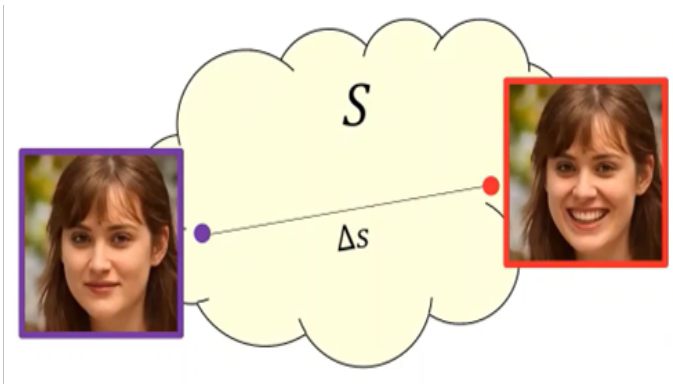
2.3 Global Direction

The global direction method is generalized for any text prompt unlike the other two approaches and it transforms the textual direction in clip space into a direction in style space to generate the offset.

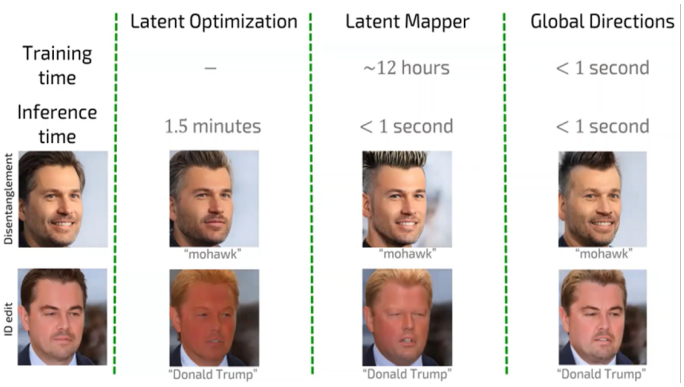


The goal is to find the global direction in style space such that by traversing along this direction the target attribute is modified for an arbitrary image. The above example explains the global direction for "gray hair". It can be observed that for all the 3 instances of gray hair, the same direction is applied and hence this approach is referred to as Global direction.
A direction star space denoted by △s, has a corresponding direction in clip space denoted by △i.
The △i is generated by applying the image encoder to a pair of before and after images. Now, this direction should agree with the direction in the cliff space defined by the text and denoted by △t. We use the two text prompts to define the direction in clip space.

In the below example, smile on the face varies while traversing in opposite direction.



2.4 Comparison of The Approaches



- Training a mapper for a new prompt takes several hours while finding a direction is immediate.
- The inference using the mapper and global direction is immediate while optimization takes more than a minute.
- The disentanglement of the global direction edit is the best thanks to the careful construction of the direction in the style space.

- The optimization in the mapper approach succeeds in manipulating identity while the global direction is not good enough for that.

3 PROBLEM DESCRIPTION

Since the Latent Mapper technique showed promising results when we factored in both disentanglement and Identity Editing, we decided to go forward with exploring and implementing this method in further detail. We want to explore the effect of the dataset's ethnicity distribution on the Latent Mapper model and analyze the training bias. The question we raised was: "What if we train a dataset with a majority of Brown/Asian faces instead of white faces? Will the ethnic bias in the training dataset affect the style transformations, which are independent of ethnicity? Eg: The "Surprised" style on the brown input face should not ideally change the color or innate features of the individual's face and should only add the "Surprised" emotion to the face. By Ethnicity bias, we mean that the generated image should retain the innate ethnic features of the individual. Innate ethnicity features of the face should not be affected by the ethnicity distribution present in the dataset. So, a person looking "Caucasian", should not look "Asian" if we have test on a model trained on an Asian dataset. So for that, we went on to create a new dataset that contains a majority of Brown/Asian faces and test the model's performance and bias. Evaluation of fairness is aimed at both the models: 1) The pre-trained model, which is trained on CelebA-HQ (white majority) dataset, and 2) New model trained on our dataset (Brown/Asian).

4 DATASET

StyleCLIP was trained on the CelebA-HQ dataset. The CelebA-HQ dataset is a high-resolution face dataset with over 30,000 images of celebrity faces. Like many face datasets, the CelebA-HQ dataset is known to have a significant racial bias. It is one of the largest and most diverse face datasets publicly available, with over 30,000 high-quality images of human faces, but it has a majority of caucasian faces. In this project, we investigate whether training a generative model on a dataset with a majority of Brown/Asian faces instead of Caucasian faces affects the style transformations of the model. Hence the objective of our project was to explore the impact of ethnic bias in training datasets on the performance of generative models for style transfer.

To get the most optimal dataset for our objective, we first searched for the Brown/Asian image dataset but did not find any with good clarity. Meanwhile, we found UTKFace dataset consisting of Brown and Asian faces majorly. So, we cleaned the UTKFace dataset (consisting of 20k faces) by filtering out Caucasian faces. After filtering the dataset, we removed any blurred or poorly identified images, resulting in a final dataset of 10,200 images for our study. The images removed were the ones, which did not have faces clearly visible by them, according to the DLIB library.

5 METHODOLGY OF OUR WORK

The latent factor model takes the most time to train against all the techniques mentioned. This is because the latent network is trained for each text prompt and it is not a generalized technique. Each text prompt for 50k steps and batch size of 2 takes up to 10 hours to train. Due to this computation and time-intensive nature of the training, we selected only a few text prompts for our analysis.

Using the e4e model, we first generated and saved the latents of every image for training a Latent mapper. To validate our results, we visualized the same latents with the StyleGAN2 decoder to confirm that the encoder (e4e) performed equally well on the new dataset as it does on the CelebA-HQ dataset.

We went ahead with the following texts:

(1) Angry
(2) Curly Hair
(3) Surprised
(4) Taylor Swift
(5) Mohawk

The loss function for the latent mapper consists of many different loss functions, added together linearly. The different losses that we considered are given below:

(a) Clip Loss: Computes the cosine distance between the manipulated image and the given text.
$L_{CLIP}(w,t) = D_{CLIP}(G(w + M_t(w)), t)$
(b) ID Loss: Computes the identity loss using ArcFace.
$L_{ID}(w) = 1 - < R(G(ws)), R(G(w)) >$
(c) L2 Loss: This is a L2 regularization term that makes sure our model doesn't get too biased.
$L_2(w) = \lambda_{L2}||M_t(w)||_2$

Hence the Combined Loss equation that we get is:
$L(w) = L_{CLIP}(w,t) + \lambda_{L2}||M_t(w)||_2 + \lambda_{ID}L_{ID}(w)$

Since we wanted to check if this model introduced any racial or ethnic biases into the modifications, we used two models for finding the inferences.

Pre-trained model - StyleCLIP's Pretrained Model that is trained on the FFHQ dataset + others.
Our model - The model we trained on our collated (UTK) brown faces dataset.

6 RESULTS

(1) Comparison of image modifications for the Text Prompt- "Surprised" using pre-trained model.

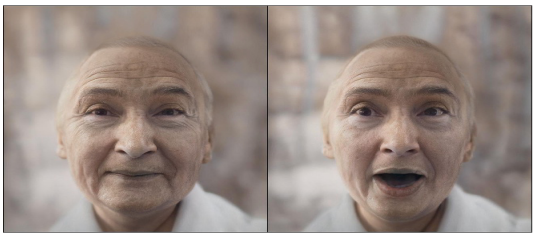


Figure 1: Curated UTK dataset

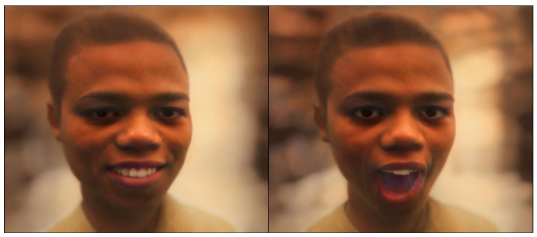


Figure 2: CelebA-HQ dataset

(2) Comparison of image modifications for the Text Prompt-"Surprised" using our model.

Figure 3: Curated UTK dataset

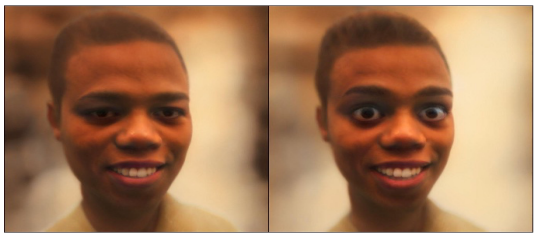


Figure 4: CelebA-HQ dataset

(3) Comparison of image modifications for the Text Prompt-"Curly Hair" using pre-trained model.

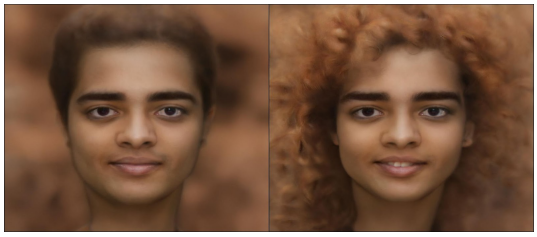


Figure 5: Curated UTK dataset

Figure 6: CelebA-HQ dataset

(4) Comparison of image modifications for the Text Prompt-"Curly hair" using our model.

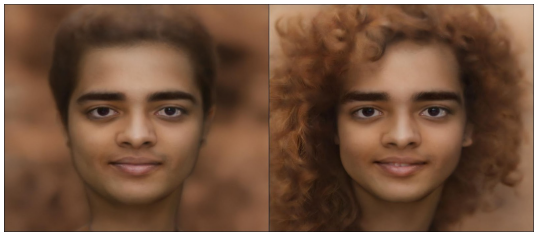


Figure 7: Curated UTK dataset

Figure 8: CelebA-HQ dataset

(5) Comparison of image modifications for the Text Prompt-"Angry" using pre-trained model.

Figure 9: Curated UTK dataset

Figure 10: CelebA-HQ dataset

(6) Comparison of image modifications for the Text Promt-"Angry" using our model.

Figure 11: Curated UTK dataset

Figure 12: CelebA-HQ dataset

(7) Comparison of image modifications for the Text Prompt-"Mohawk" using pre-trained model.

Figure 13: Curated UTK dataset

Figure 14: CelebA-HQ dataset

(8) Comparison of image modifications for the Text Prompt-"Mohawk" using our model.

Figure 15: Curated UTK dataset

Figure 16: CelebA-HQ dataset

(9) Comparison of image modifications for the Text Prompt-"Taylor Swift" using pre-trained model.

Figure 17: Curated UTK dataset

Figure 18: CelebA-HQ dataset

(10) Comparison of image modifications for the Text Promt-"Taylor Swift" using our model.

Figure 19: Curated UTK dataset

Figure 20: CelebA-HQ dataset

7 CONCLUSION

As per our analysis, we observe that prompts that focus on hairstyles ("Mohawk Hairstyle", "Curly Hair") had a negligible difference in the generated images on both pre-trained model and UTK dataset-trained models on both the datasets. As these prompts require hairstyle changes, which are nowhere related to facial features, we do not observe tonality changes.

But, in the cases of prompts that dictated changes in the expression of faces ("surprised", "angry"), our model trained on the UTK dataset performed comparatively better on the UTK dataset. We suspect that's because we trained on the UTK dataset, it allowed it to understand features like eyes, nose, mouth, etc., in a better way and allowed it to depict them w'ell.

However, we did observe the most discernible differences for the "Taylor Swift" text prompt. Since this prompt is neither a physical attribute nor an expression, but rather a whole other person. The two models interpreted it differently and very interestingly. The pretrained models seemed to change the person very definitively to look like Taylor Swift and thereby showing a prominent identity loss. Whereas our model changed only a few features of the individual to make it "taylor Swift-esque" but still maintain the ethnic integrity of the subject.

Hence, we conclude that the StyleCLIP model doesn't show ethnic biases for physical attributes or expressions, but it does vary in the extent of transformation when the text prompt is a person.

8 REFERENCES

(1) Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski: "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery"
(2) UTK Dataset: https://archive.org/details/UTKFace
(3) ArcFace model: https://github.com/peteryuX/arcface-tf2