Sketch Creation by Selecting Facial Features using CLIP guided StyleGAN3

Nikhil Jawade, MSc AI with ML, ID: 210987533

School of Electronics Engineering and Computer Science,
Queen Mary University of London, UK
ec211246@qmul.ac.uk

Abstract. In recent years, StyleGAN (a Generative Adversarial Network) and its variants have established themselves as the image generators of convincing human portraits. But all the generated faces are non-existing humans. In this study, an attempt has been made to use the ability of StyleGAN of generating a lifelike portrait to develop a real face guided by human input. In this work, Contrastive-Language-Image-Pretraining (CLIP) is used in order to develop a text-driven interface for StyleGAN image manipulation. First, the method used to generate a face by showing user options of facial features and taking their input is described, and its results are presented. Later, which approach worked, and which didn't are discussed. Finally, future improvements are suggested to this amateur effort.

1 Introduction

It is intriguing to look at the impact Generative Adversarial Network (GAN) has created since its introduction in 2014 [2]. Various academia has tried and was successful in improving the network, which revolutionized image synthesis. One of them is a Style-Based Generator Architecture for Generative Adversarial Networks or StyleGAN, which is motivated by style transfer literature with a re-designed generator architecture in a way that exposes novel ways to control the image synthesis process [3]. Not only was this new architecture introduced but a new dataset of human faces (Flickr-Faces-HQ, FFHQ) was also presented in [3].

In recent studies, it has been shown that Vision-Language models can be paired with generative models to provide a simple and intuitive text-driven interface for image generation and manipulation [6]. In this study, CLIP has been leveraged for the sole purpose.

This study aims to generate a photo-realistic image of an existing person via this architecture and human input, freeing human to verbally explain the facial characteristics of the person’s face being generated/sketched. A technique of extracting images of several faces is introduced through a CLIP guided StyleGAN3 architecture. As its human input-driven creation, these images of faces are later cropped and divided into four different categories of *Forehead, Pair of Eyes, Nose and Mouth*. The program first presents 10 choices of each feature to the human user, and then the user is asked to select one image from each of those categories based on the user's interpretation of the face to be created. Later these images of facial features are synthesized into an image of a face.

2 Background

AI in Visual Arts has picked-up talk in the last decade and has seen significant developments in the domain. Let's shed a light on some relative work in the same domain as this project and what role this background takes in shaping this project.

2.1 Related Work

There are notable mentions in the related work including a site named 'thispersondoesnotexist.com' [9] which went live in early 2019 and has engendered talks about it not only in the scientific community but also in media. DALL-E by OpenAI is another significant mention of the relative work with its state-of-the-art text-to-image generation [10].

The above relative work inspired the idea of attempting to make the face of a real human being to form an impact of AI on society to tackle situations in which a person's sketch needs to be generated, but the user (also a person) has seen that face and could not verbally explain the facial features to the sketch artist. Such situations are criminal face generation or face generation of parents of an infant who is lost.

3 System Description

In the creation of the system suitable for this project, there were several choices of Neural Nets. Some of them were tested to align them with the required configuration, which is further discussed in section 5 Discussion.

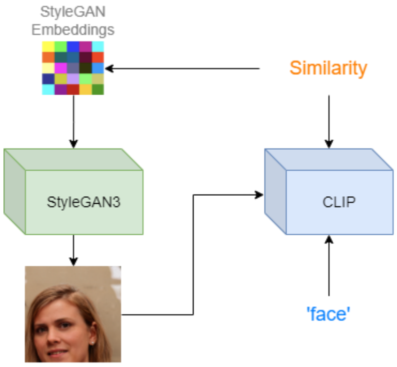


Fig. 1: Image Generation Architecture

The above architecture represents the approach taken to generate faces from prompts using CLIP and StyleGAN3. The sole purpose of CLIP was to guide an image through a text prompt which would offer more control over output generation.

StyleGAN turned out as an obvious choice because of two major reasons. One, it is trained over the FFHQ (Flickr-Faces-HQ) dataset consisting of 70,000 high-quality images of human faces having considerable variation in terms of age, ethnicity and facial accessories which aligns with the aim of this project. Second, previous related works showed promising results with this architecture.

StyleGAN3 has several significant advantages over its predecessors StyleGAN and StyleGAN2. The new StyleGAN3 generator matches StyleGAN2 in terms of FID [7] meaning it doesn't really make a huge difference when it comes to the context of this project. However, StyleGAN3 differ dramatically in its internal representation compared to StyleGAN2 and is fully equivariant to translation and rotation even at the subpixel scale [7]. The new network was able to resolve the disturbing issue of 'texture sticking' which was seen in its predecessor. The following figure shows the underlying architecture of StyleGAN3:

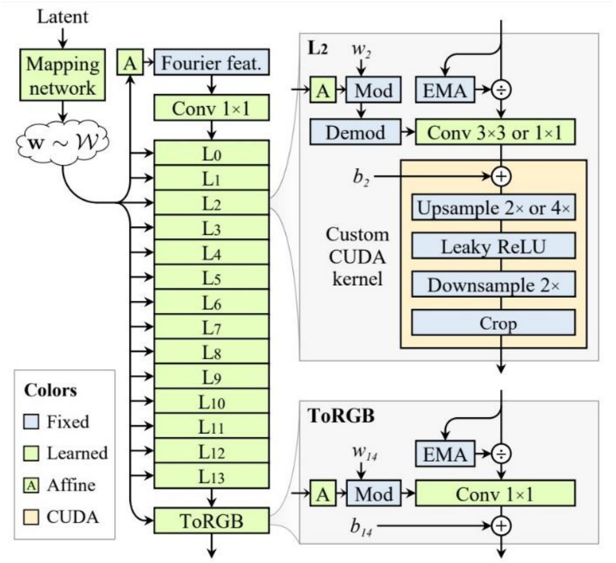


Fig. 2: StyleGAN3 Architecture

4 Experiments and Results

4.1 Experiments:

Generation of images of *Facial Features* using CLIP guided StyleGAN architecture did not return promising results (See 5 Discussion). Another huge aspect of the project is showing *Facial Features* as a choice for the human user. To match this configuration, generating several images initially and later cropping those images based on specific dimensions into 4 different parts (*Forehead, Pair of Eyes, Nose and Mouth*) approach seemed possible. Here is an example of how the image was cropped:

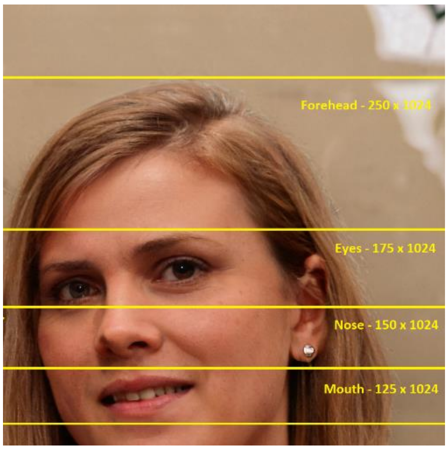
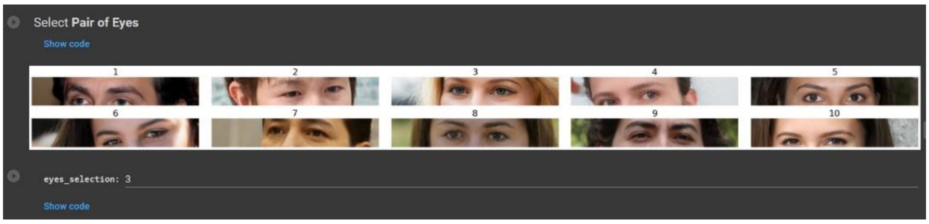


Fig. 3: Image Dimensions

Guidance of CLIP through a prompt offered stability in generating images that share similar dimensions. The prompt was set to ‘face’, and output was generated over 200+ different seeds to come up with images of faces sharing the same alignment.

After cropping images, they are divided into decided categories of features, and posed as a choice to the human user as follows:


Select **Pair of Eyes**

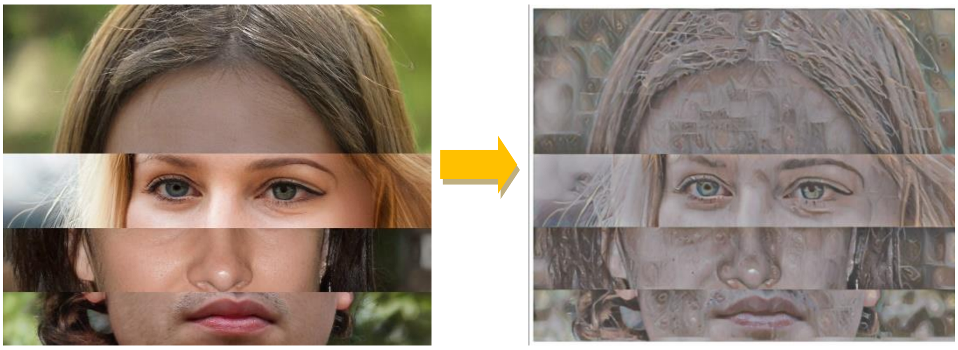
eyes_selection: 3

Fig. 4: Image Grid shown to user for feature selection

For all four features, a grid of ten images is shown to the user, and just below that, input is taken based on the number shown over the top of every image.

4.2 Results:

After combining the different feature image input from the user, the output looks like a disjointed face as it could be a mixture of different skin tones, hair colours, etc. To make the image more coherent, it is passed through itself with a simple and quick Neural Style Transfer from *the tensorflow_hub* package. Here's what the final output looks like:



(A) (B)

Fig. 5: (A) shows initial image generated after user input and (B) shows image after Neural Style Transfer

5 Discussion

Finding a use-case that will benefit the day-to-day life of people from a Computational Creativity project was the moto towards the selection of this project. One of the use-cases in mind was face sketch for forensics.

One of the works closer to this project is, Generated Photos [11], in which they offered several editable options on a StyleGAN generated portrait. But it does not deal with the randomness of image generation. Hence, creating an exact face becomes a tedious task. Specific facial portrait generation through AI is a big ask, and there were several existing NNs that were tested for this project which is discussed in the following section.

5.1 Failed Approaches:

When I started experimenting with the pre-trained GANs, the aim was to generate photo-realistic images of facial features. CLIP guided VQGAN results were quite interesting for the prompt 'eye'. Take a look at the following image:

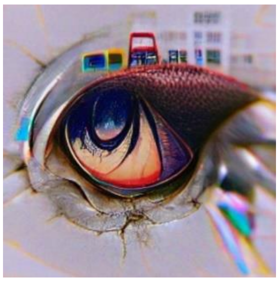
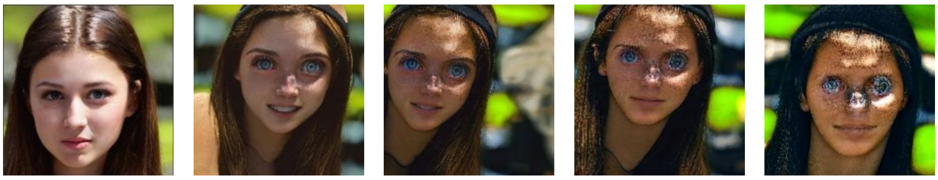


Fig. 6: VQGAN+CLIP with prompt ‘eye’ after 1000 epochs

This image is generated with VQGAN's ImageNet model trained over the dataset of 14,197,122 annotated images according to the WordNet hierarchy [4]. The appearance of the eye looks more artistic than humanlike, which did not align with the aim of the project. Moreover, it took a significant number of iterations to create a discernable image of an eye while the project demands at least 10 of those to be shown as the choice to the human user.

Another failed approach was to generate facial features with the same system used for this project. CLIP guided StyleGAN3 seemed like the best possible option to create images of features with its FFHQ model. However, outputs over different prompts for an 'eye' such as 'pair of eyes', 'close image of an eye', 'photo-realistic rendering of an eye', and more, resembled a similar pattern that is in the first epoch, a face would be generated and as it progresses the eyes keep getting highlighted giving the face a Halloween look as shown in following images:



Epoch 1 Epoch 50 Epoch 100 Epoch 150 Epoch 200

Fig. 7: StyleGAN3+CLIP with prompt ‘a pair of human eyes’

This led to the approach of cropping the generated image of the face into 4 different sections in vertical order.

6 Conclusions and Future Work

Picking apart all the relevant features of the human face and recomposing them in a way that is coherent, was the biggest challenge towards the aim of this project. The outputs are not up to the standard of being categorised as a face as this was a simple approach towards a relatively bigger aim toward societal benefits.

Moreover, this work requires significant improvements in the system used as well as its architecture. A significant number of features were not considered during this study, including ears, jaw, the colour of eyes, facial hair and its colour, etc. Additionally, adding emotions and accessories such as glasses, earrings and facial tattoos should also be given a thought in future work.

One of the recent studies at NVIDIA [8] demonstrated CLIP guided text-driven method that allows shifting a generative model to new domains, without having to collect even a single image. This approach with the right amount of data and treating every facial feature as a domain may lead to some interesting results.

References

1. Kugel, Peter. Artificial Intelligence and Visual Art. Leonardo. 15. 10.2307/1574391, 1982
2. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets” in *Advances in neural information processing systems*, 2014.
3. Tero Karras, Samuli Laine and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. ArXiv, *arXiv:1812.04948 [cs.NE]*, 2018.
4. Patrick Esser, Robin Rombach and Björn Ommer. “Taming Transformers for High-Resolution Image Synthesis”. ArXiv, *arXiv:2012.09841 [cs.CV]*, 2020.
5. Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or and Dani Lischinski. “StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery”. ArXiv, *arXiv:2103.17249 [cs.CV]*, 2021
6. Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik and Daniel Cohen-Or. “StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators”. ArXiv, *arXiv:2108.00946 [cs.CV]*, 2021
7. Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen and Timo Aila. “Alias-Free Generative Adversarial Networks”. ArXiv, *arXiv:2106.12423 [cs.CV]*, 2021
8. Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik and Daniel Cohen-Or. "StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators". ArXiv, *arXiv:2108.00946 [cs.CV]*, 2021.
9. This-Person-Does-Not-Exist 2019, accessed 25 March 2022, <https://thispersondoesnotexist.com/>
10. OpenAI 2022, accessed 1 April 2022, <https://openai.com/dall-e-2/>
11. Generated Media Inc. 2022, accessed 28 March 2022, <https://generated.photos/face-generator>