Multifaceted neural self-portrait

Harvard AC215: Advanced Practical Data Science

Dongyun Kim, Vasco Meerman

1. Background

With the advent of Artificial Intelligence, it has changed the world in different ways such as an autonomous car in computer vision and language translation in natural language processing. However, AI is only limited for those who can know how to code, have statistical knowledge and mathematics background even though it has a lot of potential to be used in many fields. In particular, Generative Adversarial Network (GAN) has shown interesting features such as latent space and feature extraction, and skilled artists or affordable artists who have in-house engineering teams take advantage of GANs as a tool to help their artworks or create their own artworks, for example, Refik Anadol while normal artists have not had opportunities to take advantage of novel technologies so far.

Figure 1. Refik Anadol, Quantum Memories, 2021

In this project, we want to open AI gates for those who are marginalized due to the technology gap and help those who are in the blind spot in technology. The application is a new kind of AI-powered creative tool for artists, designers, and creators. No need to have an AI background.

2. Project statement

Painting real people is called a portrait. The most important thing in a portrait is its recognizability. This means that when one looks at the painting, one should be able to recognize the person who is the model and this is not just about the similarity in appearance. To be a good portrait, the person's character, knowledge, and achievements must be directly or indirectly revealed. However, portraits are not always realistic. Because a portrait is drawn by order, the request of the client is actively reflected, and the artist's interpretation and perspective of the character are inevitably buried.

Today, although the medium in which portraits are drawn has changed from analog to digital, the essence of portraits remains the same; It reveals one’s identity. Countless selfies posted on Social Networking Service are intentionally manipulated and reproduced in various ways and selfies began to

replace the role of portraits. People spend numerous time modifying their selfie to make it look better and, following this trend, many SNS support built-in retouching functions in their applications. However, this modification could help add decoration but couldn't create natural changes such as posture, facial expression.

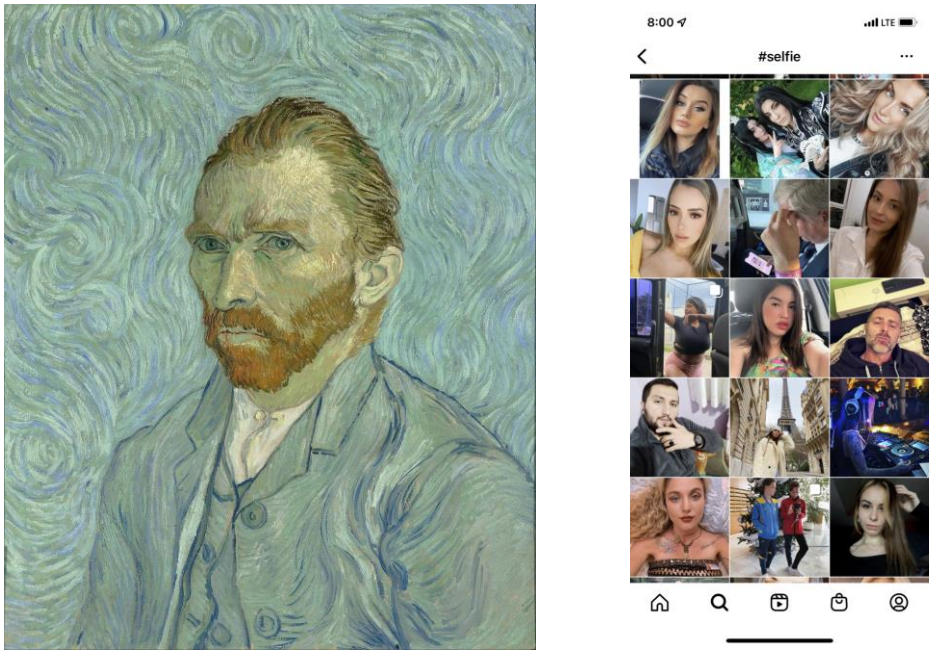


Figure 2. Vincent van Gogh, Self-Portrait, 1889 (left) and Instagram hashtag search result, #selfie (right)

Adjusting an image is easy, such as changing color and blurring boundaries with accessible image edit functions, but modifying the essence of an image (facial expression, emotion, etc.) is not easy. The small and large features of the face harmonize together to form a person, and when one characteristic disappears, this change makes it difficult to recognize the person.

Modifying an image while maintaining the facial features of a person is challenging, but if we can extract and capture these features, it might be possible to generate an image which shows different emotional expressions but can be recognized as the same person.

3. Explanation of base model

3. 1. StlyeGAN2 (https://github.com/NVlabs/stylegan2)

An alternative generator architecture for generative adversarial networks, borrowing from style transfer literature. The new architecture leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis. The new generator improves the state-of-the-art in terms of traditional distribution quality metrics, leads to demonstrably better interpolation properties, and also better disentangles the latent factors of variation. To quantify interpolation quality and disentanglement, we propose two new, automated methods that are applicable to any generator architecture. Finally, we introduce a new, highly varied and high-quality dataset of human faces.

Main feature of styleGAN is that it can capture different sizes of features from a training image set, learn the features and create latent space based on the features. The dimension of the latent vector is 512 so it helps the model store a lot of feature information in latent space. The outstanding performance on feature extraction is well aligned with our project motivation, but current styleGAN model only supports the decoder part so that we can only get randomly generated images, but couldn't

get latent vectors, which is necessary to manipulate the features of the input image.

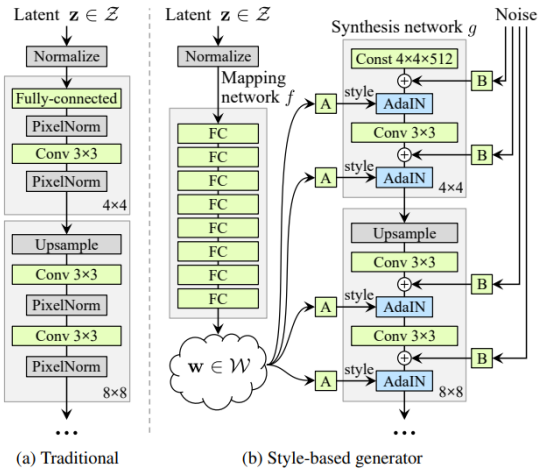


Figure 3. StyleGAN architecture

3. 2. pixel2style2pixel (https://github.com/eladrich/pixel2style2pixel)

We present a generic image-to-image translation framework, pixel2style2pixel (pSp). Our pSp framework is based on a novel encoder network that directly generates a series of style vectors which are fed into a pretrained StyleGAN generator, forming the extended W+ latent space. We first show that our encoder can directly embed real images into W+, with no additional optimization. Next, we propose utilizing our encoder to directly solve image-to-image translation tasks, defining them as encoding problems from some input domain into the latent domain. By deviating from the standard "invert first, edit later" methodology used with previous StyleGAN encoders, our approach can handle a variety of tasks even when the input image is not represented in the StyleGAN domain. We show that solving translation tasks through StyleGAN significantly simplifies the training process, as no adversary is required, has better support for solving tasks without pixel-to-pixel correspondence, and inherently supports multi-modal synthesis via the resampling of styles. Finally, we demonstrate the potential of our framework on a variety of facial image-to-image translation tasks, even when compared to state-of-the-art solutions designed specifically for a single task, and further show that it can be extended beyond the human facial domain.

The problem mentioned above could be resolved by pixel2style2pixel. It introduced styleGAN encoder so utilizing the model, we are able to get latent vectors of any input image.

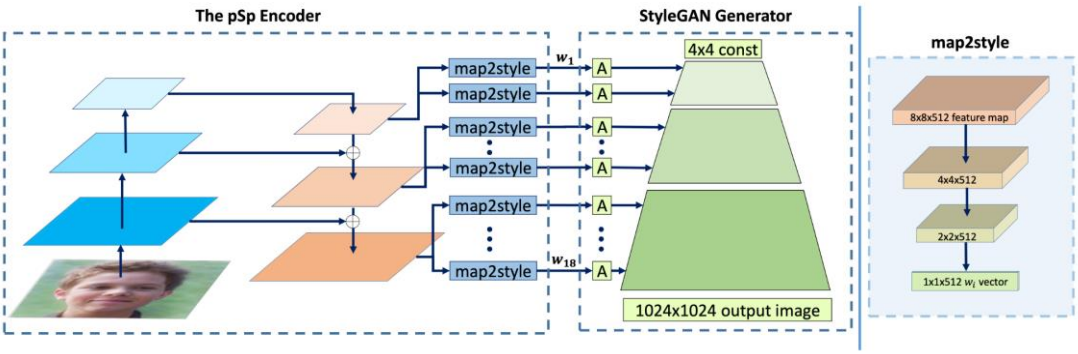


Figure 4. Pixel2style2pixel architecture

3. 3. Neural Style Transfer

The last step is to apply some artistic style to manipulated images. This allows users to add artistic values to manipulated images.

4. Baseline models trained

Basic workflow of the project is as follows.

- Users upload their selfie.
- Find the latent vectors of the user's image.
- User interface supports the latent vector manipulation.
- Manipulated vector is converted to the final image.

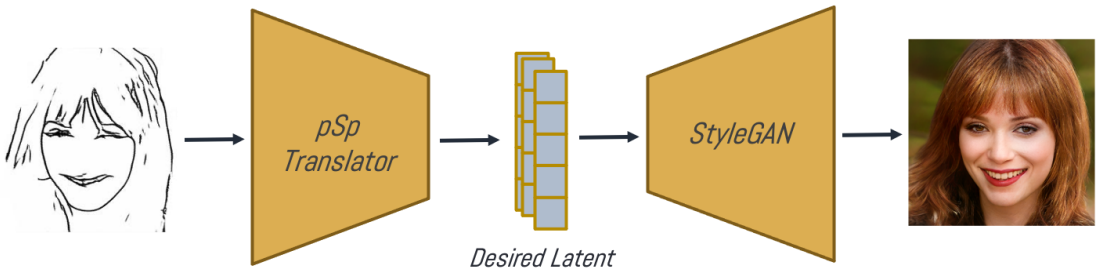


Figure 5. Image reconstruction diagram

4. 1. pixel2style2pixel

Pixel2style2pixel and styleGAN support a pre-trained network trained with FFHQ dataset (Flickr-Faces-HQ dataset at 1024×1024). The domain of the pre-trained dataset is the same as that of this project so that training process is not required. If we can narrow down our target domain such as specific race or gender, we can do additional training to get more reliable result images.

Figure 6. Original image (Left) and reconstruction image by pixel2style2pixel (Right)

4. 2. StyleGAN2

After obtaining the latent vectors of the input image, we are able to play with the vector directly. Since the latent vectors are representations of features of the image, if we manipulate the vectors to a certain direction, the converted images will show manipulated features graphically.

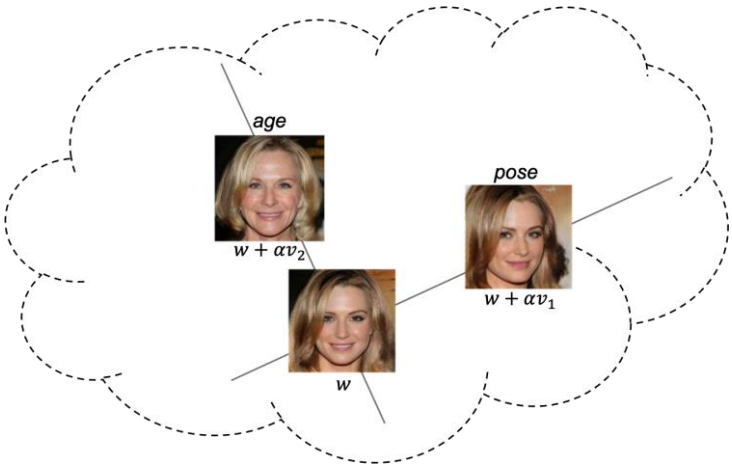


Figure 7. Manipulation of latent vectors in the latent space

The images below are examples of manipulation of the latent vectors. We move input latent vectors to the negative and positive direction of the ‘age’ feature.

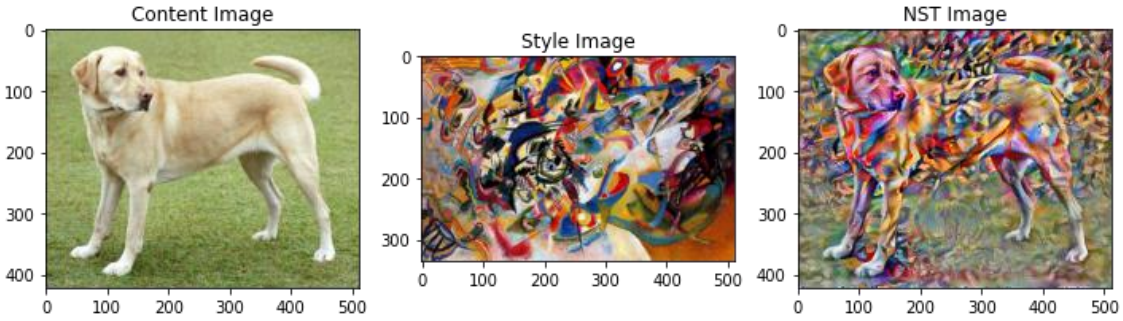
Figure 8. The image moved to the negative direction of ‘age’ feature (Left), original image (Center), The image moved to the positive direction of ‘age’ feature (Right)

4. 3 Neural Style Transfer

For our baseline NST model, we used most of the setup of the original paper called, A Neural Algorithm of Artistic Style. The model’s architecture consists of the pre-trained VGG19 without it’s classification head. From this model different layers are used for the content representation, last layers, and for the style which are multiple layers throughout the network. We use a Gram Matrix to average the correlation over the outer layers of all locations to capture the style of the original image.

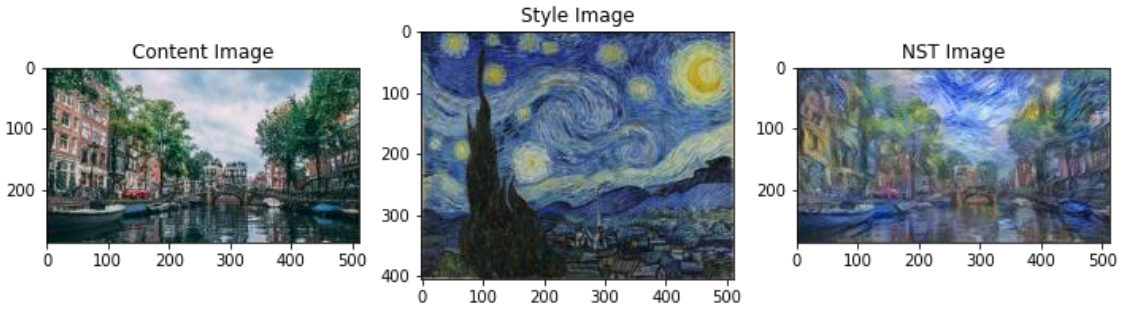
4. 3. 1. NST Experiment 1

In this experiment, we combined an image of a Yellow Labrador with Wassily Kandinsky's Composition 7.



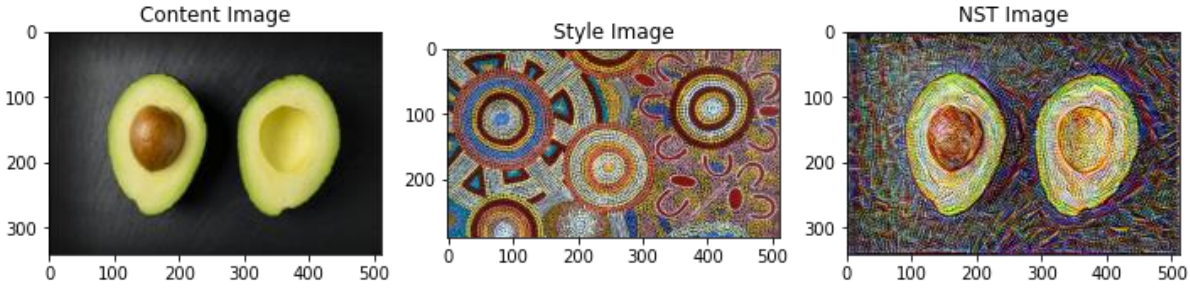
4. 3. 2. NST Experiment 2

In this experiment, we combined an image of the Amsterdam Canals with Vincent van Gogh's Starry Night painting



4. 3. 3. NST Experiment 3

In this experiment, we tried working with a different type of 'art', we used a form of art called 'dot-art' for which an object is painted in with colorful dots. We also ran this experiment to determine how well this setup would work with the styled image being an actual image compared to an image of a painting. We used a relatively simple image with similar shapes as our content image, an open avocado.
We did also play around with the hyperparameters like style weights and the Adam learning rate to get a more visually stable image, since the output also looks more like a painting then an image in some particular style.



4. 3. 4. NST Results

Our experiments showed promising results from our baseline model. We noticed that the original architecture didn't perform as when we used a real image as 'styled' input compared to an image of a painting. We played around with multiple hyper parameters, like our weighted losses and Adam optimizer learning rate.

```
[39] opt = tf.optimizers.Adam(learning_rate=0.005, beta_1=0.99, epsilon=1e-1)
```

To optimize this, use a weighted combination of the two losses to get the total loss:

```
[40] style_weight=1e-3
     content_weight=1e4
```

This unfortunately did not improve our third experiment as much as would have hoped. Also, using this setup of Neural Style Transfer requires the model to learn specifically for the content and style image, which took a few minutes on a macbook without GPU support, making it not a very effective option to use in a live application.
This all concludes that we will need to further investigate different newer Neural Style Transfer models with the following capabilities:

- Relatively short inference time
- Able to predict on data is hasn't seen before
- Optionally: A one model approach to generate different styles for an input image

5. Main functions

5. 1. Upload selfie

When users play with their own images, it tends to make them more involved in the application than using given images. This will create a strong bond between users and the application.

5. 2. Navigate latent space with features

We are now supporting 15 features, age, eye distance, eye eyebrow, eye ratio, eyes open, gender, lip ratio, mouth open, nose mouth distance, nose ratio, nose tip, pitch, roll, smile and yaw. With simple vector arithmetic, we can add and subtract input image vectors to these feature vectors and feed the result vector into styleGAN generator model so that the image with changed features could be generated.

5. 3. Apply artistic style with Style Transfer

Manipulated images can be imported to add artistic styles. It makes the image look like more artistic portrait than a real image. Also, users have another chance to modify their images.

6. Technical challenges

Pixel2style2pixel is only developed by pytorch, but pytorch requires specific GPU allocation which might cause some problems when we use GCP. Thus, we need to find some way to convert pytorch model to tensorflow. We luckily found ONNX which helps convert torch to tensorflow.

7. Future functions considered

7. 1. Automatic crop and align input image

It is important to put well-aligned input images to get good latent vectors because styleGAN is sensitive to the alignment of images. However, when users upload their selfie, the images may not be properly aligned. Thus, to support good performance of the model, additional automatic crop and alignment of input image will be helpful.