SAPIENZA
UNIVERSITÀ DI ROMA

Style-based GANs – Generating Realistic Artificial Statue Faces

GARBA MARIAM : 1871871
GIORGIA PIERNOLI: 1648511

Prof. Fiora Pirri
Vision and Perception

July 29, 2019

Contents

1.0 Introduction

Generative Adversarial Networks (GAN) are a relatively new concept in Machine Learning, introduced for the first time in 2014. Their goal is to synthesize artificial samples, such as images, that are indistinguishable from authentic images. A common example of a GAN application is to generate artificial face images by learning from a dataset of celebrity faces. While GAN images became more realistic over time, one of their main challenges is controlling their output, i.e. changing specific features such pose, face shape and hair style in an image of a face.

A paper by NVIDIA, A Style-Based Generator Architecture for GANs (StyleGAN), presents a novel model which addresses this challenge. StyleGAN generates the artificial image gradually, starting from a very low resolution and continuing to a high resolution (1024× 1024). By modifying the input of each level separately, it controls the visual features that are expressed in that level, from coarse features (pose, face shape) to fine details (hair color), without affecting other levels. This technique not only allows for a better understanding of the generated output, but also produces state-of-the-art results – high-res images that look more authentic than previously generated images.

In this project, StyleGAN is used to generate artificial statue faces by learning from a dataset of Statue faces which was gathered from Pinterest.

2.0 Background

The basic components of every GAN are two neural networks – a generator that synthesizes new samples from scratch, and a discriminator that takes samples from both the training data and the generator's output and predicts if they are "real" or "fake" . The generator input is a random vector (noise) and therefore its initial output is also noise. Over time, as it receives feedback from the discriminator, it learns to synthesize more "realistic" images. The discriminator also improves over time by comparing generated samples with real samples, making it harder for the generator to deceive it.

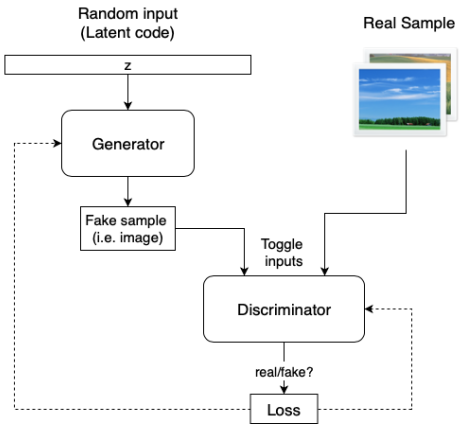

Figure 1: GANs overview

Researchers had trouble generating high-quality large images (e.g. 1024× 1024) until 2018, when NVIDIA first tackles the challenge with ProGAN. The key innovation of ProGAN is the progressive training – it starts by training the generator and the discriminator with a very low resolution image (e.g. 4× 4) and adds a higher resolution layer every time.

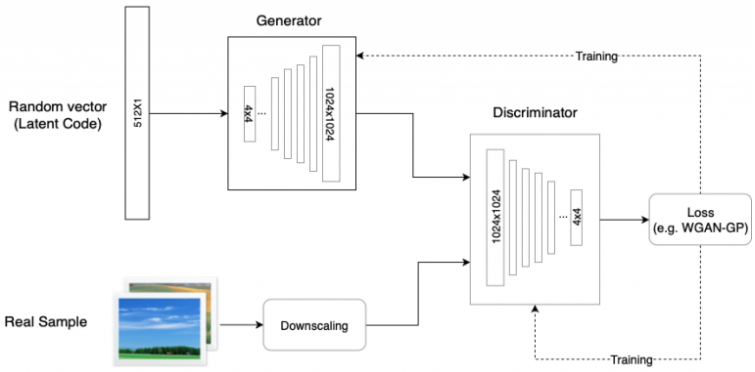

Figure 2: ProGAN overview

ProGAN generates high-quality images but, as in most models, its ability to control specific features of the generated image is very limited. In other words, the features are entangled and therefore attempting to tweak the input, even a bit, usually affects multiple features at the same time. A good analogy for that would be genes, in which changing a single gene might affect multiple traits.

3.0 Methodology and Implementation

The StyleGAN is an upgraded version of ProGAN's image generator, with a focus on the generator network. It was observed that a potential benefit of the ProGAN progressive layers is their ability to control different visual features of the image, if utilized properly. The lower layer (and the resolution), the coarser the features it affects. These features were divided into three types:

1) Coarse – resolution of up to 8^2 – affects pose, general hair style, face shape, etc
2) Middle – resolution of 16^2 to 32^2 – affects finer facial features, hair style, eyes open/closed, etc.
3) Fine – resolution of 64^2 to 1024^2 – affects color scheme (eye, hair and skin) and micro features.

The new generator includes several additions to the ProGAN's generators:

3.1 Mapping Network

The Mapping Network's goal is to encode the input vector into an intermediate vector whose different elements control different visual features. This is a non-trivial process since the ability to control visual features with the input vector is limited, as it must follow the probability density of the training data. The Mapping Network consists of 8 fully connected layers and its output w is of the same size as the input layer (512×1).

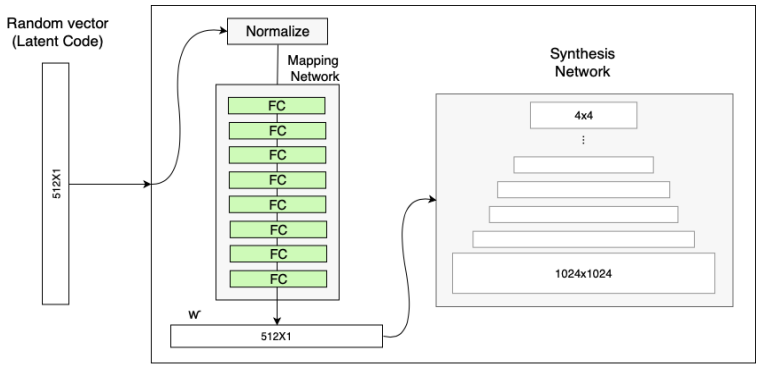

Figure 3: The generator with the Mapping Network (in addition to the ProGAN synthesis network)

3.2 Style Modules (AdaIN)

The AdaIN (Adaptive Instance Normalization) module transfers the encoded information w, created by the Mapping Network, into the generated image. The module is added to each resolution level of the Synthesis Network and defines the visual expression of the features in that level:

1) Each channel of the convolution layer output is first normalized to make sure the scaling and shifting of step 3 have the expected effect.
2) The intermediate vector w is transformed using another fully-connected layer (marked as A) into a scale and bias for each channel.
3) The scale and bias vectors shift each channel of the convolution output, thereby defining the importance of each filter in the convolution. This tuning translates the information from w to a visual representation.

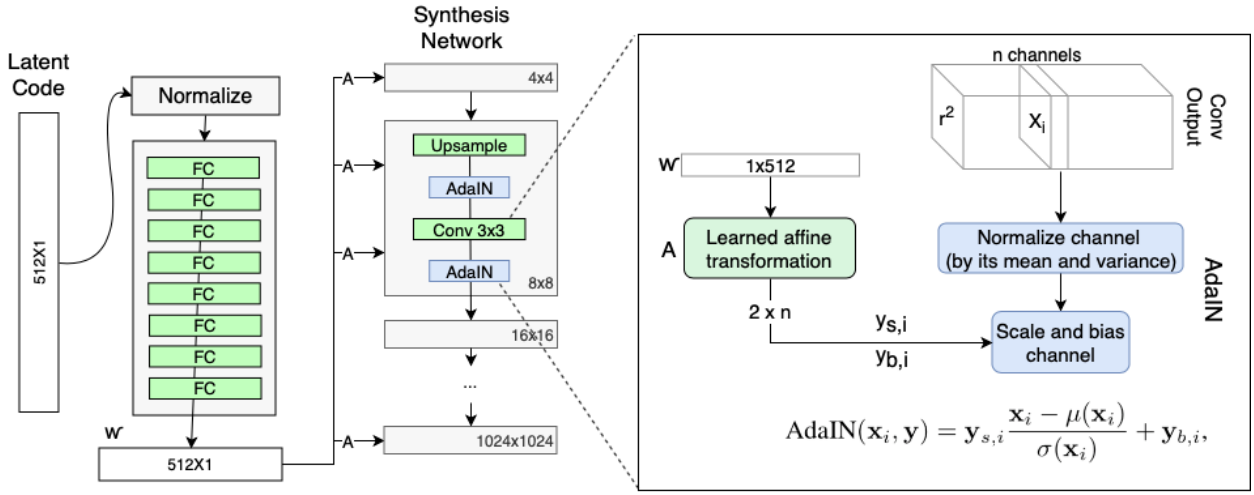

Figure 4: The generator's Adaptive Instance Normalization (AdaIN)

3.3 Removing traditional input

Most models, and ProGAN among them, use the random input to create the initial image of the generator (i.e. the input of the 4× 4 level). In StyleGAN, the initial input are replaced by constant values.

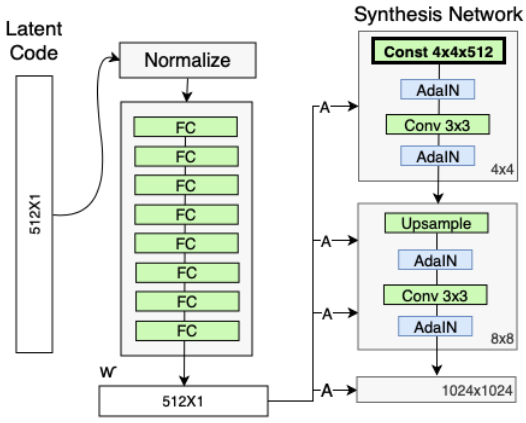

Figure 5: The Synthesis Network input is replaced with a constant input

3.4 Stochastic variation

The noise in StyleGAN is added in a similar way to the AdaIN mechanism – A scaled noise is added to each channel before the AdaIN module and changes a bit the visual expression of the features (freckles, hairs etc.) of the resolution level it operates on.

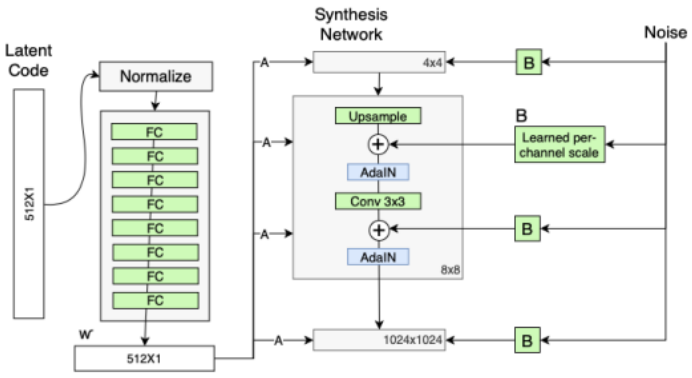

Figure 6: Adding scaled noise to each resolution level of the synthesis network

3.5 Style mixing

The StyleGAN generator uses the intermediate vector in each level of the synthesis network, which might cause the network to learn that levels are correlated. To reduce the correlation, the model randomly selects two input vectors and generates the intermediate vector w for them. It then trains some of the levels with the first and switches (in a random point) to the other to train the rest of the levels. The random switch ensures that the network won't learn and rely on a correlation between levels.

3.6 Truncation trick in W

To avoid generating poor images, StyleGAN truncates the intermediate vector w, forcing it to stay close to the "average" intermediate vector. After training the model, an "average" w_{avg} is produced by selecting many random inputs; generating their intermediate vectors with the mapping network; and calculating the mean of these vectors. When generating new images, instead of using Mapping Network output directly, w is transformed into w_{new}=w_{avg}+$(w - w_{avg})$, where the value of defines how far the image can be from the "average" image (and how diverse the output can be).

3.7 Fine-tuning

This step is significant for the model performance. Additional improvement is made by updating several network hyperparameters, such as training duration and loss function, and replacing the up/downscaling from nearest neighbors to bilinear sampling.

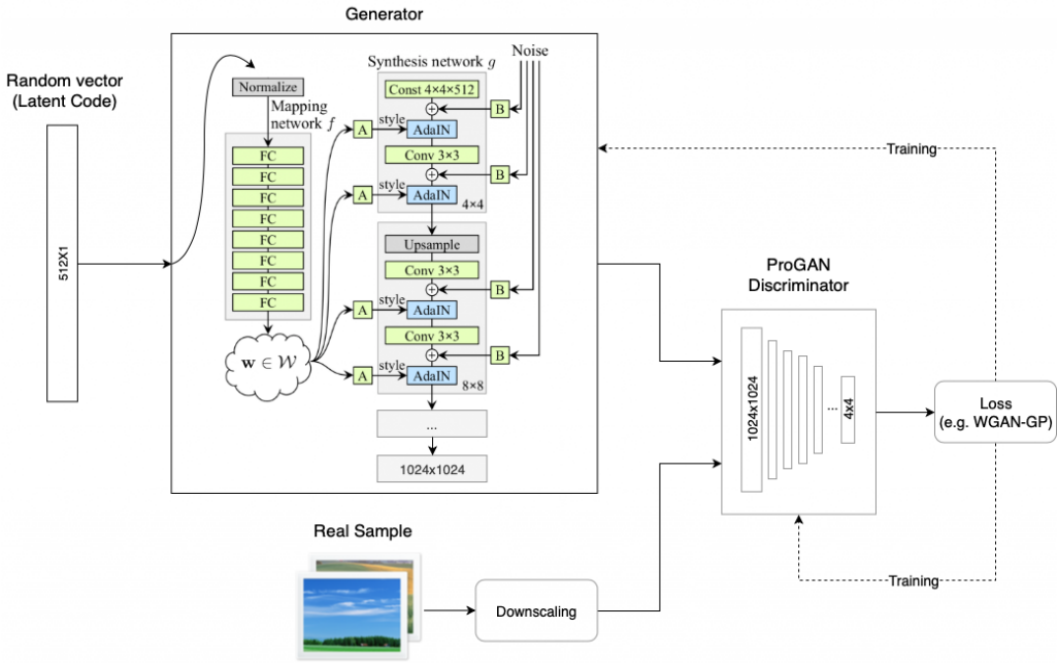

Figure 7: StyleGAN Architecture

3.8 Implementation Details

StyleGAN was trained on the dataset of Statue faces which was gathered manually from Pinterest and it consists of images of 106 portrait statues. We build upon the TensorFlow implementation of StyleGAN for anime portraits which we inherited most of the training details. In particular, we used the same discriminator architecture, resolution dependent minibatch sizes and exponential moving average of the generator. Extra modifications were made on the hyperparameters to improve the overall result quality.

Also for the dataset, we scraped Statue faces from pinterest and performed some preprocessing techniques (cropping and resizing) on them in order to make them suitable and rich for the training process. This dataset was chosen in order to test if StyleGAN would perform very well on complex structures different from human faces.

4.0 Results

This study presents results on the dataset of Statue faces. The figure below shows an excerpt of Statue portraits that was used during the training process as well as the Statue portraits that were generated by the StyleGAN after training.

Figure 8: Training Data Excerpts

Figure 9: Generated Statues (Cherry Picked)

Using all the methodology of StyleGAN that has been explained previously, we can see that the generator did a good job at generating believable fake Statue portraits.

5.0 Conclusion

StyleGAN is a groundbreaking technique that not only produces high-quality and realistic images but also allows for superior control and understanding of generated images, making it even easier than before to generate believable fake images. The techniques presented in StyleGAN, especially the Mapping Network and the Adaptive Normalization (AdaIN), will likely be the basis for many future innovations in GANs.

References

https://www.scihive.org/paper/1812.04948

https://arxiv.org/abs/1406.2661

https://arxiv.org/abs/1710.10196