Neural Network Theorem
神经网络理论

Shanghai Jiaotong University – College of Software Engineering

Final Project

Human Face Stylizing

人脸图片风格化

李陈豪 120037910063 lee_vius@sjtu.edu.cn

June 8, 2021

Contents

1 Introduction

In recent years, there are many works that study the fields of facial expression transfer and face stylizing, which have great potential in various application of industries, including films, commercial and entertainment, etc.

Since the publish of GAN[1] network, for generative adversarial network, in 2014, many researchers have proposed GAN-based methods for image generation. Cycle-GAN[3] and StyleGAN[2] have made great work in generating images based on needs of users. Cycle-GAN takes advantage of characteristics of GAN network, and tries to transfer images from one domain to another, which makes image stylizing possible. StyleGAN, though with style as its name, cannot complete domain transfer for data. It focuses more on the features of image and tries to let the GAN network have more control on features of different scales in generated images, and the ability to control detailed generation of images is referred to ”style”. Though Cycle-GAN can complete the transfer between data of different domains, it requires large amounts of data to train a fine-tuned style interpreter. For facial images, it is hard to get huge amounts of stylized images, which may require artists to take long time to create. For example, sources of animation-type faces and Disney-type faces are limited, and it’s quite hard to create a database containing huge amounts of data for training a style transfer model.

In this work, I will take advantage of the great feature controlling ability of Style-GAN network. Based on this characteristic, I’ll propose a model-blending method based on transfer learning to achieve the target of image stylizing.

2 Related Work

In recent years, many works have made significant progress in image stylizing. Many GAN-based[1] networks have achieved great effects in generating images. GAN, for generative adversarial network, proposed a novel training idea. The network can play a great role in data generation and augmentation. It assumes that two agents, generator and discriminator, are gaming with each other. The generator will generate data from nothing, and give that data to the discriminator. The discriminator need to strictly judge whether the data received belongs to the database we have. Through this way, the generator can finally generate data that confuses the discriminator, which means GAN network can generate data belonging to distribution of database.

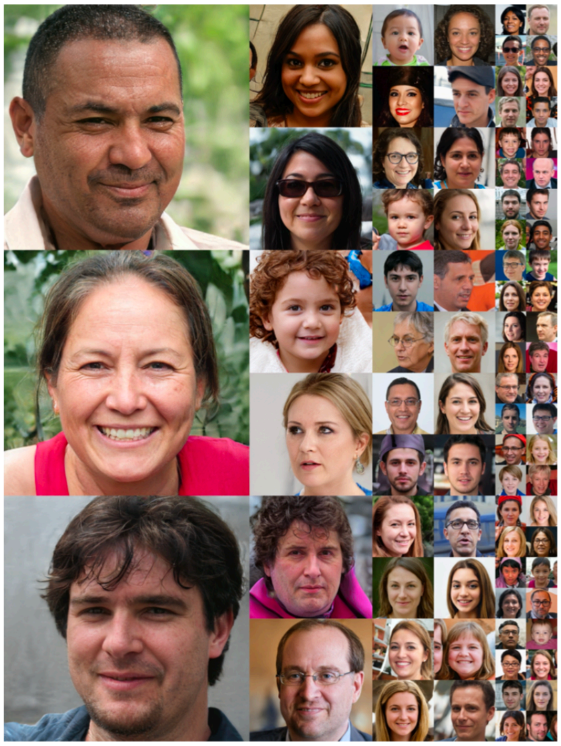

Figure 1: StyleGAN-generated Human Faces

Many GAN-based networks in recent years can generate data belonging to certain domain. StyleGAN[2] makes even further study on how to control features of generated images, which can finally generate high-resolution images with rich details. Based on GAN network structure, StyleGAN weaken the concept of input layer in the generator. It inputs the latent code to the convolution layers of different scales as a control method. The training results in a model that can generate different features at convolution layers of different scales. StyleGAN can generate images of high quality even for high-resolution images, whose distribution is very close to real data.

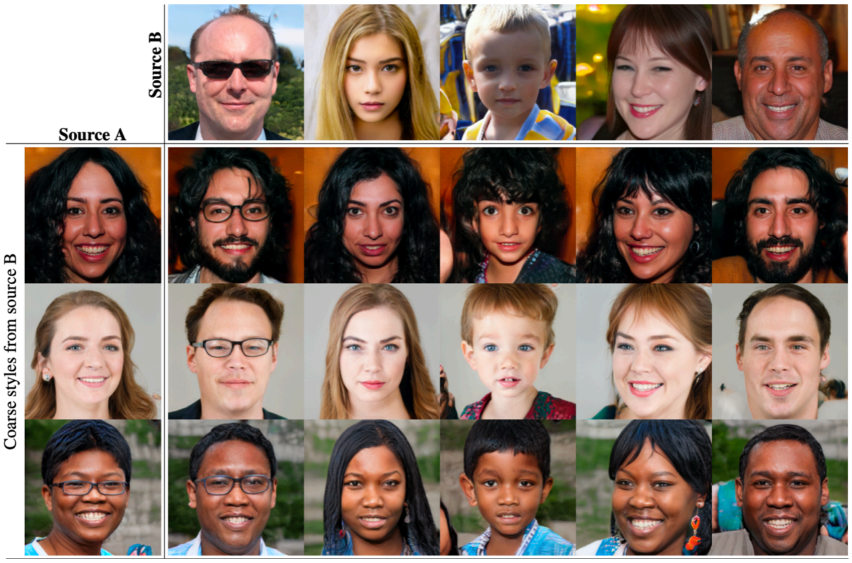

Figure 2: StyleGAN: Style-mixed Human Faces

Taking advantage of characteristics of convolution layer at different scale, a StyleGAN generator can control various attributes of images generated. Combined with extra training, users can control detailed information on the generated images, including sex information, age, or even certain facial expression. By blending the latent codes corresponding to two faces, StyleGAN can generate a face containing both features of them, which is referred to ”Style” in the paper. However, StyleGAN itself cannot compete the work of stylizing, which means transferring an image to another domain. However, transfer learning, unlike training from nothing, will have lower cost for StyleGAN, and doesn't require very large amount of data.

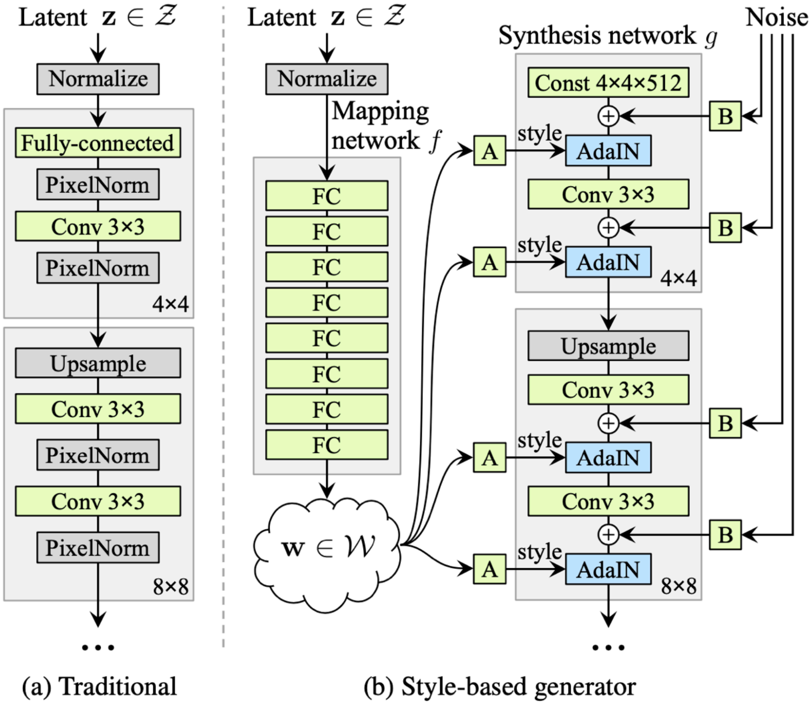

Figure 3: StyleGAN Network Structure

Cycle-GAN is a method to accomplish domain transfer, which can transfer an image to another style. The significant point of the model is to propose a cycling network structure. When the model transfer an image to another domain, and then transfer the generated image back, we should ensure the final image is the same as the original image. During this process, Cycle-GAN proposes the cycle-consistency loss to improve the performance. Though Cycle-GAN can accomplish image style transfer, it doesn't have the elegant characteristics of StyleGAN, for example, having detailed control on the generated images.

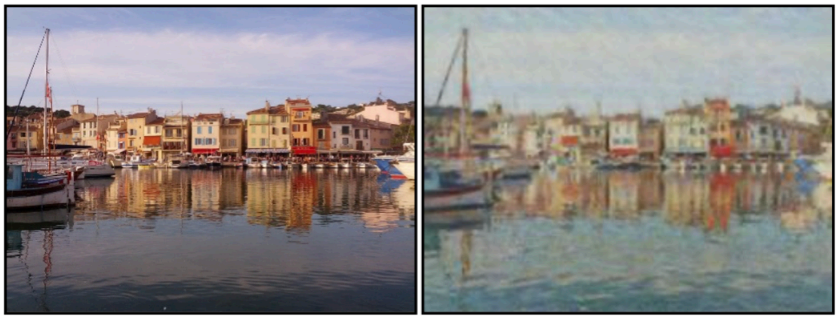

Figure 4: Cycle-GAN applies style transfer

Based on the work above, my method will try to conserve the characteristics of StyleGAN, and meanwhile implementing the image stylizing. I'll take advantage of transfer learning based on fine-tuned models and then blend the models to get a style-mixed StyleGAN image generator.

3 Method

The overall method are dividing into two parts, which are transfer learning and model blending. Transfer learning is working on a fine-tuned StyleGAN2[2] model and small set of data from another domain, such as animated faces or Disney-cartoon faces.

3.1 Fine-tuned learning

The first problem we need to solve is the data set for training. The advantage of transfer learning is the low requirement of data amount and training time. As we are training the facial image data, and the only difference is that we want to generate faces of certain styles instead of real human faces, we can assume the distributions of two domains of data are close to each other. Thus, transfer learning can be expected to be quite easy and fast.

This project will mainly focus on two types of data, including animated data and Disney-type faces. Animated faces data are relatively easy to get, which will be given in the Section Results. For Disney-type faces, we need to construct data set ourselves. I collected about 26 Disney films, and used scripts to crop the clips we need for training. The time-consuming part is that it requires human to provide the start and end frames, together with the bounding box in the film to extract the clips. After collecting data, I passed about 350 images to StyleGAN2 model for training, based on the pretrained real face generator trained on FFHQ dataset.

Consider the training time, workload and hardware, I only trained for a 256x256 resolution StyleGAN2 generator. StyleGAN2 has its own training schedule, and I set the training starting point from 10000 timg in the code.

3.2 Model blending

After transfer learning, we should get a StyleGAN2 model that can randomly generate images belonging to certain domains. Now we need to apply our metric to enable the model to stylize real human face.

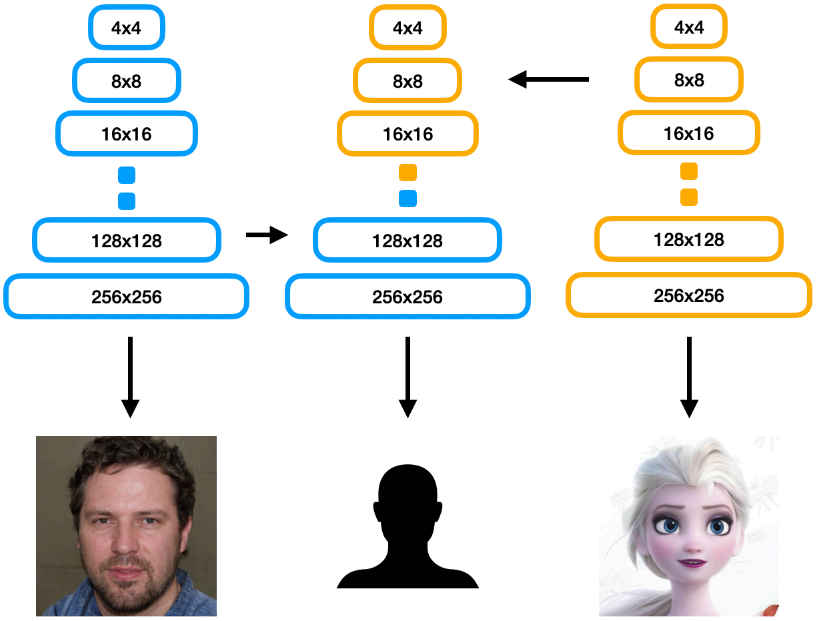

Figure 5: Model Blending Method

As described in StyleGAN[3], the generator model contains convolution layers of different scales from $2^2 \times 2^2$ to $2^8 \times 2^8$. The lower resolution layers are more likely to control detailed patterns, such as local expression, shapes, etc, in the generated images, while the higher resolution layers are more likely to generate macro information like texture and color of generated images. The blending metric is applying on the GAN model itself. As the transferred model is trained based on the real face model, we can expect that the weights in the network of these two models have continuous relation, and we can swap certain layers of one model into the other. The output model is expected to be able to generate style-mixed images containing both features of real faces and animated faces.

4 Results

This section will show the transfer learning results and model blending results. The target styles used in this project are animated faces and Disney-type faces.

4.1 Training details

As the training of StyleGAN2 model has a quite high requirement of GPU, I did the training on Colab platform, which provides Tesla T4 GPU with RAM of 16GB. It basically supports full schema of training StyleGAN2 model. Considering training time cost, I used pretrained model which is trained on FFHQ dataset and output images of resolution of 256×256. The transfer training setting is the same as original StyleGAN2 model, and the only difference is that I set the training schema starting from 10000 timg, which will fit better for a pretrained model. The traing time is about 2 hours for 24 timg.

4.2 Animated Face style

The animated face images come from Danbooru2020, a large scale rowdsourced and tagged anime illustration dataset. I randomly picked 500 images as training dataset for StyleGAN transfer training.

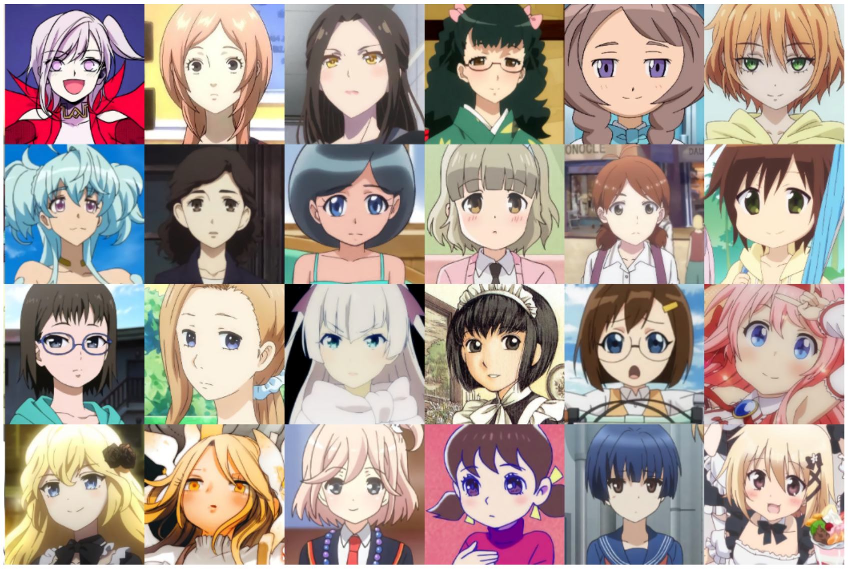

Figure 6: Animated Face training data

After transfer learning, StyleGAN2 can automatically generate random animated faces. However, as the amounts of data is small, the network will easily go overfitting once training for long epoch, and in my case, the generated images seem over-fit after training 480 epochs.

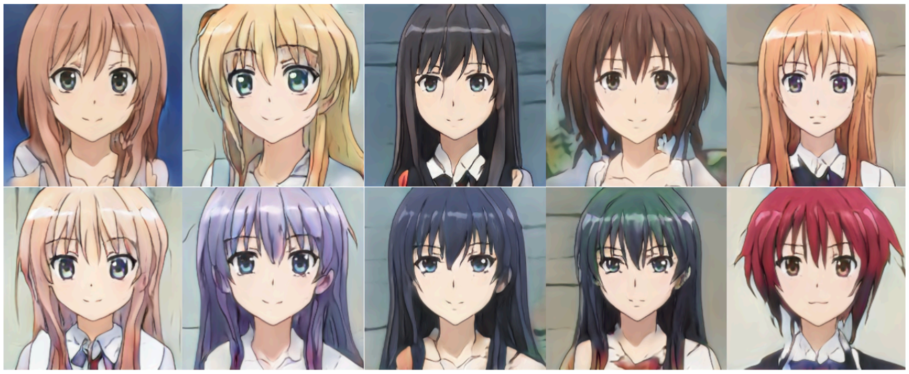

Figure 7: Random Animated Faces Generated after 160epoch transfer learning

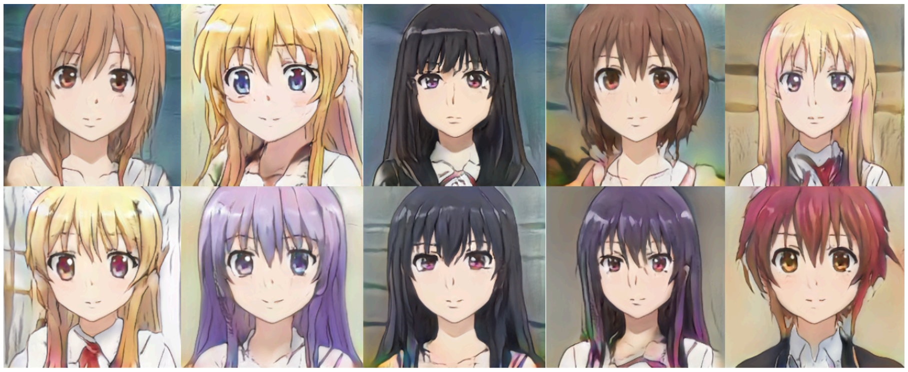

Figure 8: Random Animated Faces Generated after 480epoch transfer learning

Now we have transfer learnt models and the original real face generator, and we can apply model blending method to the two models to get a real-face stylizing generator. I blended the low resolution layers of real-face generator with the high resolution layers of anim-face generator. The swapping point is at 64×64 layer, and the results are shown below.

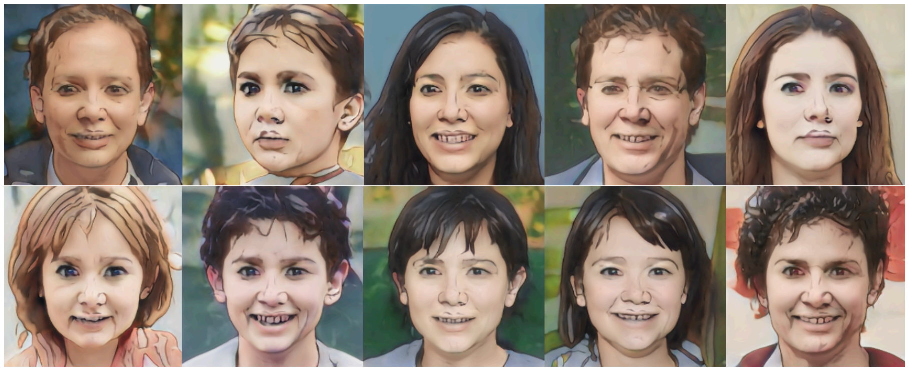

Figure 9: Blend with 160 epoch transferred model

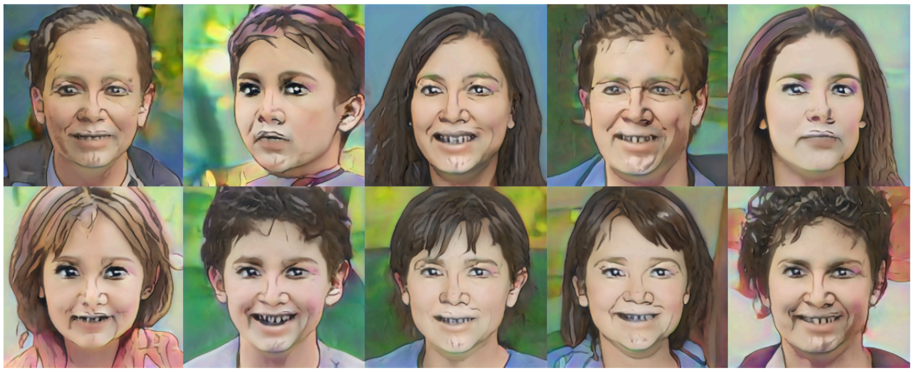

Figure 10: Blend with 480 epoch transferred model

The original real faces generated by the original model is shown below.

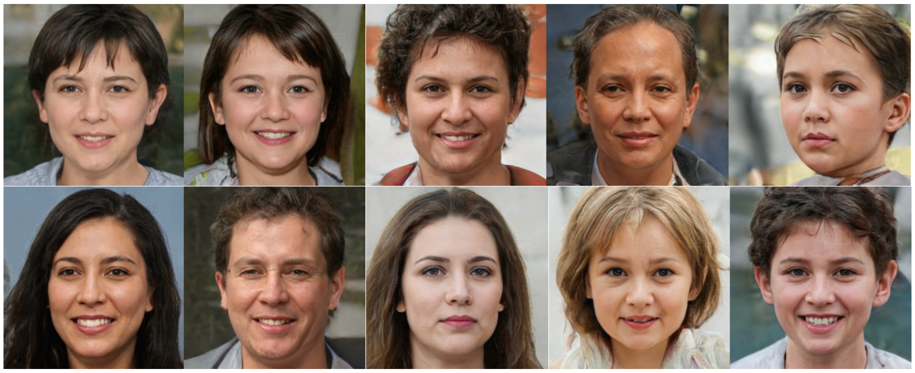

Figure 11: Real faces generated by original model

From the results, we can see that as the training time increases, the blended faces become worse. As we can observe from the dataset we are using, the animated faces are much simpler than the real faces, and they don't have as much rich details as human faces have. As the training time increases, the StyleGAN network trained itself in a way that weakens its power of generating detailed information, which finally leads to overfitting.

Model blending method can enable the model to try combing features of images of both styles. In this experiment, the blended model can generates faces with animated-style color, texture and lines, while it can reserve the shape of real faces.

4.3 Disney-type Face style

The Disney-type faces are cropped by myself from Disney films. I temporarily collected 297 clips where Disney characters making facial expressions, and from them I randomly picked certain frames as the image dataset for StyleGAN2 transfer training.

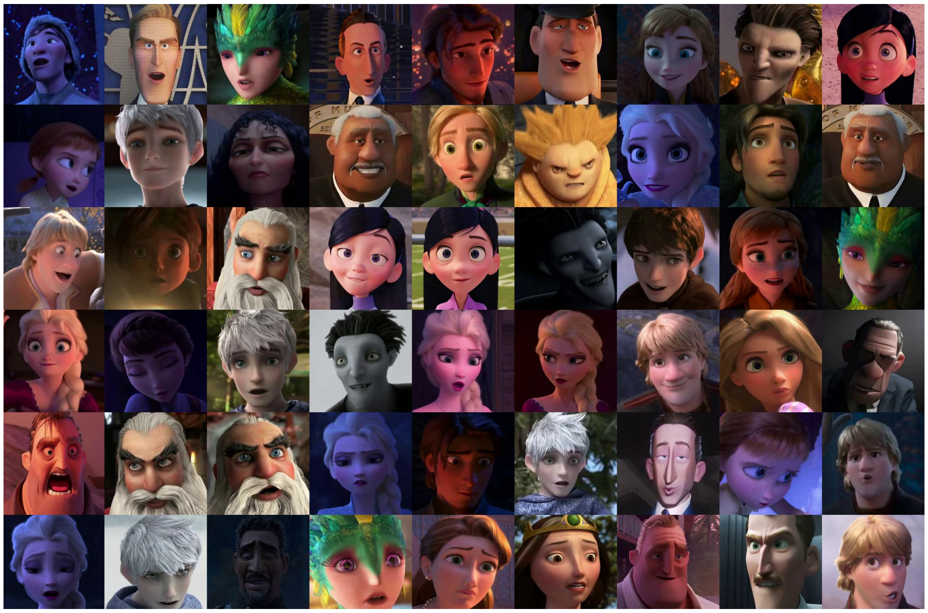

Figure 12: Example Disney-type Face training data

After transfer learning, StyleGAN2 can automatically generate random cartoon faces. As the network will easily go overfitting once training for long epoch, I set the training time relative short in this experiment, which is from 24 epochs from 96 epochs. Observing from the results, the generated images seem to go overfitting after 48 epochs.

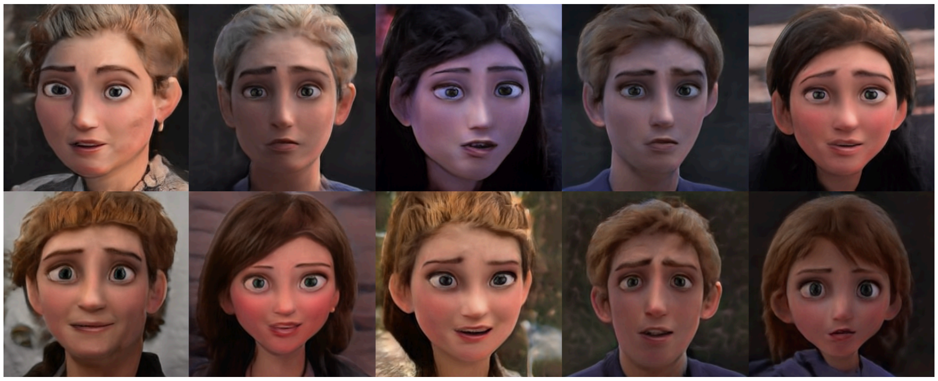

Figure 13: Example Cartoon Faces generated training 24 epochs

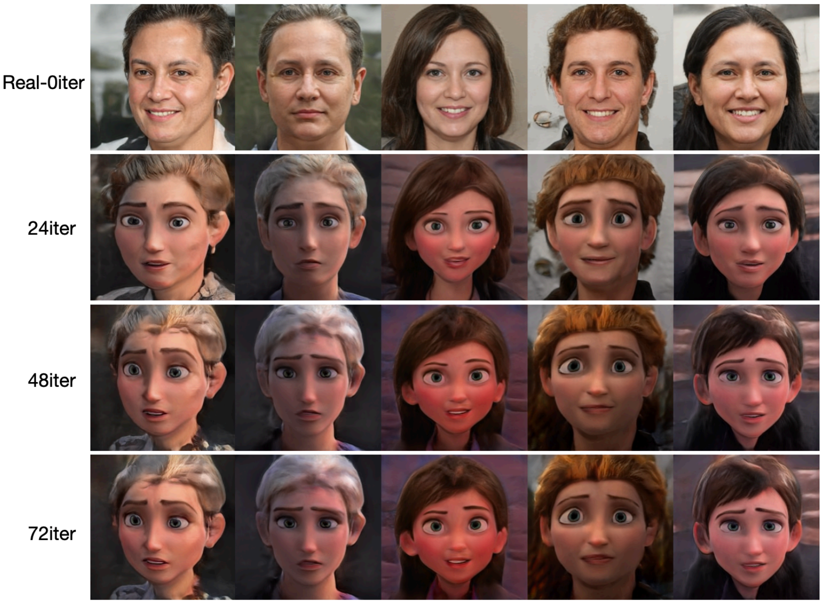

Figure 14: Example Cartoon Faces generated training 24 epochs

The real faces generated by the original model are like the example below.

Figure 15: Example Real Faces generated by original model

Then we can apply model blending method to the results. In this experiment, I blend the models from 24 epochs to 72 epochs with original real face model, and tried to swap at different convolution layers from 16 × 16 to 64 × 64. The figure below shows the results generated the blend models. The real faces generated by the original model are like the example below.

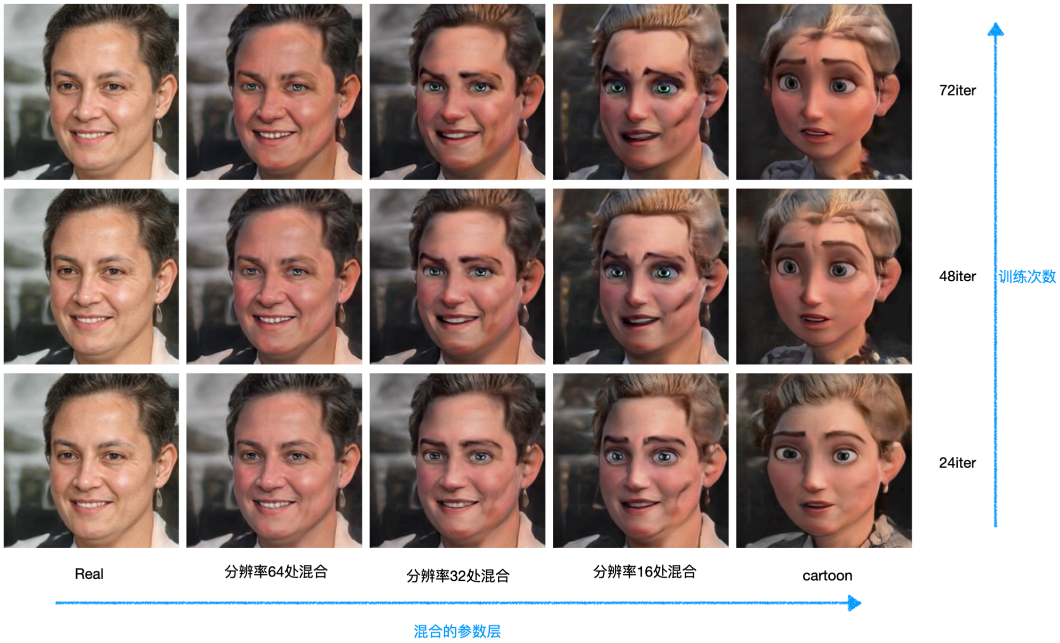

Figure 16: Example Real Faces generated by original model

From the results, we can see that the blend models can generate faces with features of cartoon faces magically when the blending point is at 64 and 32 resolution layers. The more cartoon model is blended, the more likely the generated faces to be close to a cartoon face. As we can see at blending point of 32 resolution, the blended model can already give a face that have large eyes and cartoon-like color and texture.

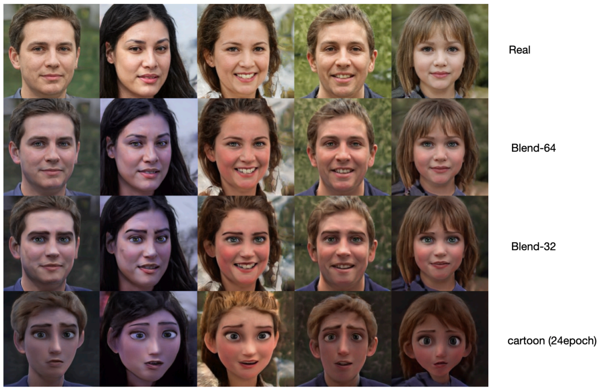

Figure 17: More blended examples: Model blended with 24 epoch trained cartoon model

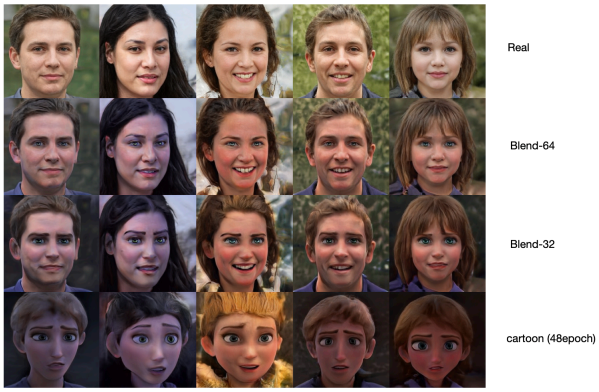

Figure 18: More blended examples: Model blended with 48 epoch trained cartoon model

References

[1] Goodfellow, I. J. , et al. ”Generative Adversarial Nets.” MIT Press(2014).

[2] Karras, T. , S. Laine , and T. Aila . ”A Style-Based Generator Architecture for Generative Adversarial Networks.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2019.

[3] Zhu, J. Y. , et al. ”Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.” IEEE (2017).