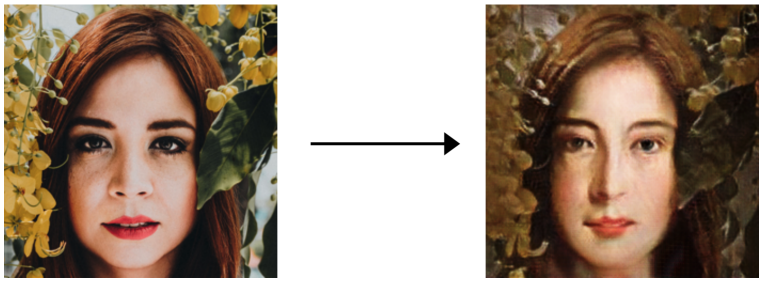
CS230

Transforming Portraits to Classical Paintings using CycleGAN

Phillip Kim
Stanford University
pkkim@stanford.edu

Helen Wu
Stanford University
helenwu8@stanford.edu



Abstract

The goal for this project is to modify the CycleGAN [1] model to successfully generate digital portraits, in the style of 18th Century paintings, from an input of a digital photo of a human face. We aim to generate paintings that resemble the input human face while also resembling the classical painting style. We approach this problem by introducing the novel approach of using embeddings of images generated by the FaceNet model [2] as part of our loss training. Furthermore, we experiment with incorporating the Structural Similarity Index (SSIM) [3] in our loss function.

We use a dataset of paintings from Pinterest for style and a dataset from Yonsei University for real portrait inputs. We find that while all models achieve decent results, the original CycleGAN model yields the best overall quantitative results. We also find that incorporating FaceNet embeddings as part of the model's cycle consistency loss improves the generated paintings' aesthetic quality ratings and resemblance of the original image during human qualitative evaluation.

1 Introduction

The stylization of head portraits is a long-standing challenge for non-photorealistic rendering (NPR) research community. GAN networks have recently become the favourite technique for image to image translation tasks. [4] The main drawback with this system is that it requires a large dataset of paired photos. For the scope of this problem, this is extremely hard to obtain as this would require us to have a large set of digital portraits and their paired paintings in a particular style.

CycleGANs have provided a breakthrough in unpaired image training and we propose applying CycleGANs to this problem scope to overcome some of the key issues we've mentioned. The main drawback in previous attempts at unpaired alternatives to GAN is that these require a large amount of artwork in a particular style. This issue is overcome by CycleGANs which can produce impressive results on relatively small datasets (small datasets from 200 to 1000 photos are adequate [1]). Another issue in related previous unpaired approaches is that models had difficulties in reproducing delicate high-frequency details that are important to retain fidelity of used artistc media. [4] Improving on this will be a key focal point of our project.

One novel approach we introduce to this problem scope is the use of Facenet embeddings in our loss function and training. Historically, the FaceNet [2] system is used for identity verification even when the picture depicts different

features, such as age and expression, by focusing on comparing specific key features of the face. We also propose the use of SSIM (Structural Similarity) Loss [3] which aims to match the general luminance, contrast, and structural information of images.

With art style portrait generation, it is vital that we preserve the likeness of the inputted portrait but also allow the model room to generate differences that will better preserve the character and nuance of the art style. Through employing these methods, we hope to train our model with this flexibility.

The input to our model is an digital portrait of a real human face. We will then use a modified CycleGAN to output a stylised portrait in the style of 18th Century paintings.

2 Related work

In researching different approaches for our project, we examined our key problem at hand: we wanted our model to preserve the identity of the inputted portrait yet we also wanted it to have enough flexibility to alter features in order to match the desired art style. To tackle this challenge, we drew inspiration from the FaceNet model [2] which was primarily used for identification purposes. The FaceNet model generated and compared embeddings of facial features from two inputted portraits to determine if the same individual is depicted in both images. One key problem this project tried to overcome was that two images of the same person taken at drastically different times could have large differences. For example, there could be changes to the individual's hair style, clothing, and age amongst other things that could make it harder to accurately evaluate their indentity. We believed that using these embeddings as a loss function for our model would be effective as they allow us to train for the identity of the inputted individual to be recognisable in the output without necessarily optimizing to make all facial features between the two images identical.

For a similar reason, we also drew inspiration from the CycleGAN Face-Off project [5], a case of style transfer where the facial expressions and attributes of one person could be fully transformed to another face. Though tackling a different scope, this project had similar challenges to ours. Namely, it aims to transfer certain elements from one face, the expressions from the original face, to another while preserving another set of elements, the identity of the second face. This research proposed the use of SSIM (Structural Similarity) Loss [3] which matches the luminance(l), contrast(c), and structure(s) information of the generated image and the input image. SSIM has proven in the past to be very helpful in improving the quality of image generation and we have added this approach into our methodology.

3 Dataset and Features

Our CycleGAN model requires two datasets: a set of digital photos of real human faces and collection of painted portraits from the 18th Century.

For our first dataset, we took the "Real and Fake Face Detection" dataset from Yonsei University [6] and isolated 1100 pictures of real people at 600 x 600 resolution. We based our dataset size off of the number of images use to train "Horse to Zebra" CycleGAN model [1] which shows that around 1000 images is reasonable for training and testing CycleGANs.

To obtain our second dataset, we used a web-scraper tool to scrape 900 images from Pinterest tagged under "18th Century Potraits". We then manually pre-procssed the images by removing incorrectly tagged or poor quality photos so that all images had at least 412 x 412 resolution (the standard resolution used for training images of human faces), and cropping around the face.

Once inputted into our model, all our images will be automatically cropped and downsized to 240 x 240 resolution before testing and training begins.

4 Methods

4.1 Loss Formulation

The training objective for all models is to minimize the combination of three loss functions:

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + L_{cyc}(G, F).$$

where X and Y are the two domains of given training samples, G and F are mapping functions where $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and D_Y and D_X are adversarial discriminators, where D_X aims to distinguish between images $\{x\}$ and translated images $\{F(y)\}$, respectively, and vice versa for D_Y. For both mapping functions, we have adversarial losses, expressed (in this case specifically for G) as:

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim p_{data}(y)}[\log D_Y(y)] + E_{x \sim p_{data}(x)}[\log(1 - D_Y(G(x)))],$$

where G attempts to generate images $G(x)$ to fool D_Y into being unable to distinguish between x images and $G(x)$ images. In addition, we utilize cycle consistency loss:

$$L_{cyc}(G, F) = E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] + E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1],$$

which reduces the space of possible mapping functions by minimizing the L1 distance between original images and the reconstructed images, given by $F(G(x))$ and $G(F(y))$. Modifications to this cycle consistency loss are made in approaches 4.3 and 4.4.

4.2 Vanilla CycleGAN

We train a basic CycleGAN model with our portraits and classical paintings datasets. For the generator we use 3 convolutional encoding layers, followed by ResNet-9 blocks, and 2 decoding layers, all with ReLU activation. For our discriminator we use PatchGAN, containing 6 convolutional layers with LeakyReLU activation.

4.3 SSIM Loss

We propose to modify the cycle consistency loss to not only record distance between the input image and reconstructed image, but also the SSIM [3] between the images. The new cycle loss looks like:

$$L_{cyc}(G, F) =$$
$$E_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] - SSIM(F(G(x)), x) +$$
$$E_{y \sim p_{data}(y)}[||G(F(y)) - y||_1] - SSIM(y, G(F(y))),$$

where $SSIM(x, y)$ measures the structural similarity between image x and y. A higher SSIM score indicates greater perceived quality of the reconstructed image.

4.4 FaceNet Loss

We propose to change the cycle consistency loss to measure the Euclidean distance between the embeddings for original images and reconstructed images obtained from putting them through a pretrained FaceNet model. The new cycle loss would be:

$$L_{cyc}(G, F) = ||FaceNet(F(G(x))) - FaceNet(x)||_2$$
$$+||FaceNet(G(F(y))) - FaceNet(y)||_2,$$

where $FaceNet$ represents the feature representations of the images extracted from the FaceNet model.

5 Experiments

5.1 Evaluation methods

We use two different quantitative metrics and human evaluation to evaluate our models.

IS We use Inception Score [7] which applies an Inception-v3 network pretrained on ImageNet to measure the model's ability to generate images with meaningful objects and ability to generate diverse images. The score represents the KL-divergence between the two probability distributions that represent these qualities; a greater score means that a dataset is more diverse and meaningful.

FID We use Fréchet Inception Distance [8] which compares the statistics of generated samples to real samples by computing the distance between two multivariate Gaussians pulled from the Inception-v3 pool3 later. A lower score is better as it denotes a greater similarity between real and generated samples.

Human Evaluation Besides the above quantitative evaluations, we ask human volunteers to assess the artistic quality of our generated images and to distinguish between generated portraits and real classical paintings.

5.2 Experiment Details

We train all our models with the AdamW optimizer [9], performing mini-batch gradient descent. Models are trained on 200 epochs, using linear learning rate decay. The learning rate remains consistent for the first 100 epochs and decays linearly to 0 for the next 100 epochs.

For hyperparameter tuning, models were trained on 150 epochs on a random sample of 350 images pulled from the full dataset for both style images and real images in order to obtain results more quickly.

5.3 Results

5.3.1 Quantitative Metrics

Model	FID	IS
Vanilla CycleGAN	**103.677**	3.2141
FaceNet Loss	122.247	2.6551
SSIM Loss	115.108	**3.2723**

Table 1: Quantitative evaluation results.

As presented in Table 1, we see that the Vanilla CycleGAN yields the best FID and SSIM Loss model yields the best Inception Score. In terms of FID, Vanilla CycleGAN performs significantly better than all the other models. In terms of Inception Score, the SSIM Loss model's performance is closely followed by Vanilla CycleGan, while the FaceNet Loss strays far behind. Overall, these results show that despite our modifications, the Vanilla CycleGAN model continues to perform the best in generating paintings overall based on these categories.

Learning Rate	FID	IS
α = 0.0001	125.218	**2.82074**
α = 0.0002	**122.948**	2.73021
α = 0.0004	125.735	2.76173

Table 2: Learning rate tuning results.

No. of Layers	FID	IS
3	123.957	2.57377
4	122.932	2.80197
5	**121.469**	**2.86012**

Table 3: Discriminator depth tuning results.

Presented in Table 2 and 3 are results for hyperparameter tuning experiments for the Vanilla CycleGAN model. From these experiments, we see that by increasing the number of layers of the discriminators in the CycleGAN model, the FID and IS improve incrementally.

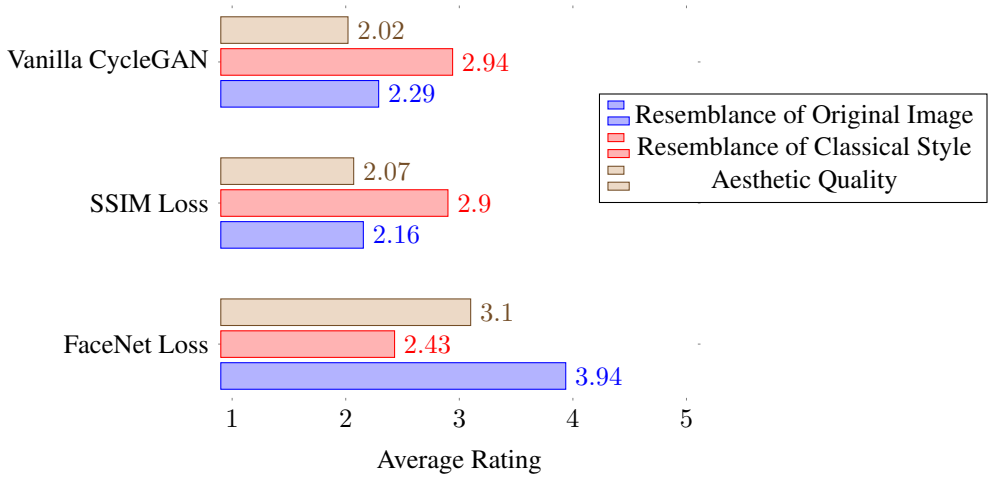


Figure 1: Average survey ratings for quality of generated paintings obtained from human volunteers.

Figure 2: Generated images sampled from test set.

5.3.2 Human evaluation results

The survey results in Figure 1 show that the FaceNet Loss model's generated paintings have the best aesthetic quality and greatest resemblance of the original image by far, while having the least resemblance of the classical style. The Vanilla CycleGAN and SSIM Loss models' generated paintings yield similar average survey ratings for each category, both performing much better than the FaceNet Loss model in terms of representing classical style.

5.4 Analysis

We see that Inception Scores are relatively low for all models. However, this is not unexpected since a set of portrait paintings would have a low diversity score with respect to the 1000 classes that the Inception network (pretrained on ImageNet) was trained to classify for. This low IS also indicates low image quality in that the probability distribution in classifying the images is not distinctly one class. This shows room for improvement on these models in producing more meaningful looking portrait paintings. Furthermore, we see that the Vanilla CycleGAN and SSIM Loss Inception Scores are significantly better than the others. We expected the IS for SSIM Loss to be high since the loss function specifically optimizes for images of better perceived quality. However, we were hoping that the FaceNet Loss model would also be able to improve upon the vanilla CycleGAN's IS by focusing on retaining important facial features of the input image. In Figure 2, we see that the samples of FaceNet Loss generated images tend to be darker and less vibrant than the images generated by the other models. This discoloration may have potentially led to loss of detail and thus a lower Inception Score.

In terms of FID, we are not surprised that the Vanilla CycleGAN retains the best score. The modifications made to the loss function in both the FaceNet Loss and the SSIM Loss models were designed to strengthen the cycle part of the model or, in other words, the CycleGAN's ability to reconstruct the original image. There were no changes made to the models in order for them to generate paintings that better match the distribution of real classical paintings. Hence, the FID would not improve given these modifications.

The survey results show that the FaceNet Loss model was most effective in retaining resemblance to the original image. This may have occurred due to the entire cycle consistency portion of the loss function being displaced with the Euclidean distance of the FaceNet embeddings. From further inspection, we found that the cycle consistency loss for the FaceNet Loss model was weighted much greater than it was in the vanilla CycleGAN model. This likely led the model to prioritize cycle consistency over the adversarial losses, meaning the model would focus more on retaining attributes of the input image rather than converting it to the style of a classical painting. This may also be why we see a low rating for resemblance of the classical style for the FaceNet model. Furthermore, this focus on resemblance of the input image may be correlated with the high rating for aesthetic quality since it would be less prone to abnormal distortions of facial features (i.e. in Figure 2 image (a), the teeth are morphed into lips in both the vanilla and SSIM model, whereas they are retained in the FaceNet image).

From a qualitative standpoint, we observe that the out-put images from the FaceNet Loss more accurately translates the details of the input images than Vanilla CycleGAN or SSIM Loss: The facial features are better preserved and there are also less irregularities with the generated image. In contrast, images generated by the Vanilla CycleGAN and SSIM Loss contain noticeable distortion to the images that causes key facial features to become deformed. As we can observe in image (d) produced by SSIM Loss, there are unpleasant textural irregularities with the woman's face.

As a result of these distortions, the images generated by FaceNet Loss are more aesthetically pleasing by comparison. However, in prioritising preserving the structural similarity to the inputted face, the FaceNet Loss outputs sacrifice the ability to accurately capture the Victorian art style. Observing images in column (a), the FaceNet output appears an almost identical copy of the input image with a color filter placed over it. Conversely, the SSIM and Vanilla images reconstructs the boy's nose to be slimmer with a high arch and also depicts a more serious demeanor – both traits which are typical in Victorian portraiture.

6 Conclusion

This project aimed to take in photos of individuals and render digital portraits in the style of 18th Century paintings that preserved the inputted identity. We used the Vanilla CycleGAN as our baseline approach and proposed novel approaches which used the FaceNet [2] and SSIM [3] models as part of our loss training.

Our dataset included classical paintings scraped from Pinterest and a dataset of real digital portraits from Yonsei university. Though our novel approaches both had their merits – in particular, the FaceNet outputs generated the greatest likeness of the inputted images and best aesthetic quality – we found that the vanilla CycleGAN yields the best overall quantitative results.

Moving forward, we plan to improve our models with further training and parameter learning strategies. We hope to experiment with weighting our modifications to the loss function differently (for both SSIM loss and FaceNet loss) to see if we can balance the trade-off of resembling the original image versus resembling the classical painting style. We also plan to experiment with incorporating the use of FaceNet embeddings in other parts of the model, including the generator architecture. We hope that our model can eventually make an impact on the problem we defined in this project.

7 Contributions

Phillip was responsible for designing and implementing the FaceNet Loss and SSIM Loss model. Phillip set up the code for training, testing, and FID and IS evaluations. For the final report, Phillip contributed to the methods, evaluation methods, results, and analysis portions.

Helen was responsible for the main literature review and determining novel approaches for the project. Helen scraped and pre-processed the datasets and also worked on the implementation of FaceNet loss. For the final report, Helen contributed to the introduction, related works, dataset and features, and analysis portions.

References

[1] `junyanz/pytorch-CycleGAN-and-pix2pix`. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[2] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[4] D. Futschik, M. Chai, C. Cao, C. Ma, A. Stoliar, S. Korolev, S. Tulyakov, M. Kučera, and D. Sýkora. Real-time patch-based stylization of portraits using generative adversarial network. In *Proceedings of the 8th ACM/Eurographics Expressive Symposium on Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*, Expressive '19, page 33–42, Goslar, DEU, 2019. Eurographics Association.

[5] Xiaohan Jin, Ye Qi, and Shangxuan Wu. Cyclegan face-off. *CoRR*, abs/1712.03451, 2017.

[6] https://www.kaggle.com/ciplab/real-and-fake-face-detection.

[7] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.