CS726: Advanced Machine Learning

Identity Aware Portrait Generation using CycleGAN

Jay Sawant - 18D070050
Yash Gadhia - 180100130
Pranav Page - 18D070020

10 April, 2022

1 Problem Statement

The aim of this project is to modify the CycleGAN model for portrait generation using human photos i.e. photo to portrait conversion. The goal is to generate portraits that preserve the facial features of the input human face while also resembling the classical painting style.

2 Related Work

Generative Adversarial Networks (GANs) [2] have achieved impressive results in image generation and representation learning [4]. GANs' success is the idea of an adversarial loss that forces the generated images to be, in principle, indistinguishable from real photos. This loss is particularly powerful for image generation tasks, as this is exactly the objective that much of computer graphics aims to optimize.
Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. However, for many tasks, paired training data will not be available. Zhu et al. [8] proposed **Unpaired Image to Image translation** which uses Cycle-GAN model which takes an image of a horse as input and translates it to a Zebra image. Motivated from this, we choose the source domain as the images of Human Faces and the target domain as portrait paintings where all the images are unpaired. However, facial content from CycleGAN cannot be well preserved because of the weak content constraint. Inspired by dual learning, Yi et al. [6] propose Dual-GAN with a similar unpaired training mechanism based on unsupervised performance
Fang et al.[1] presented an approach to translate human faces into sketches using Cycle-GAN [8]. It improves CycleGAN on photo-sketch synthesis by paying more attention to the synthesis of key facial regions, such as eyes and nose, which are important for identity recognition.

3 Datasets and Code

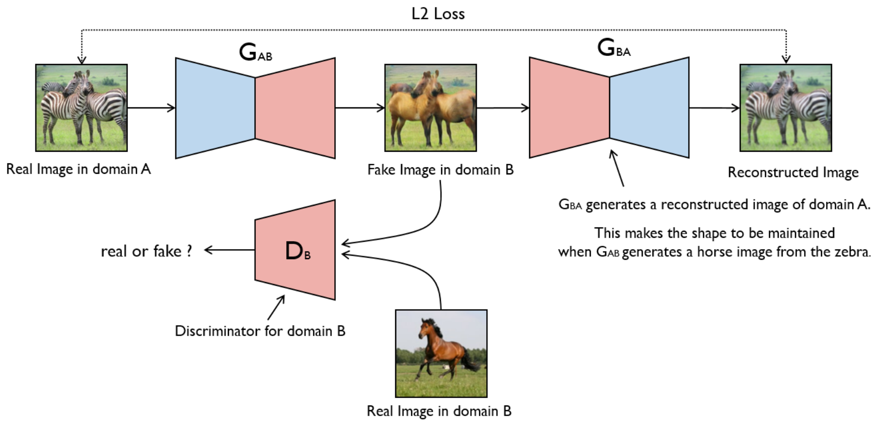
We use two datasets for the source domain and the target domain. For the source domain, we use the dataset of Human faces available on Kaggle - Human Faces consisting of 7.2k+ images useful for multiple use cases such image identifiers, classifier algorithms etc. For the target domain, we use the dataset of Kaggle - Portrait Paintings which is scrapped from the WikiArt website. This dataset consists of 5.5k+ portrait paintings for purposes like GAN training, etc. We use 1000 images from both domains for training and also create validation and test sets with 500 images each.
We are using PyTorch framework in Python for the training purposes. For the Cycle-GAN model architecture and training code, we are referring to the original PyTorch code by the authors available at https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix [8][3]
Our code is available at `https://github.com/YashGadhia/AML_Project`

4 Proposed Approach

We propose a modification of CycleGAN for portrait generation which explicitly considers the problem of recognition. More specifically, we propose to use the FaceNet model[5] to extract facial features which can then be used to guide the portrait generation network by modifying the loss function.

4.1 Cycle GAN Formulation

We denote the two domains of given training samples as X and Y, 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.



4.2 Loss Function

4.2.1 Adversarial Loss

For both the mapping functions, we have the standard adversarial GAN loss

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

where G attempts to generate images $G(x)$ to fool D_Y into being unable to distinguish between x images and $G(x)$ images. Similarly for F.

4.2.2 Cycle Consistency Loss

For each image x from domain X, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We call this forward cycle consistency. Similarly, for each image y from domain Y , G and F should also satisfy backward cycle consistency: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. All this when put together in a loss function, we get:

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

4.2.3 Proposed Perceptual Loss

We propose an additional perceptual loss that uses FaceNet features to guide the generation networks to preserve facial features between the image and its translation. For example, if we consider network G, we want the facial features between the image x, and its generated portrait G(x) to be alike. Similarly for network F. Hence the proposed loss is

$$\mathcal{L}_{\text{perceptual}}(G, F) = E_{x \sim p_{\text{data}}(x)}\left[\|FaceNet(G(x)) - FaceNet(x)\|_2^2\right] + E_{y \sim p_{\text{data}}(y)}\left[\|FaceNet(F(y)) - FaceNet(y)\|_2^2\right]$$

4.2.4 Full Objective Function

All the above losses put together we get:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda_c \mathcal{L}_{\text{cyc}}(G, F) + \lambda_p \mathcal{L}_{\text{perceptual}}(G, F)$$

5 Experiments and Results

5.1 Evaluation Metric

To evaluate the performance of the model, i.e. to compare the given human face and the generated portrait, we use the **Structural Similarity Index Measure (SSIM)** metric. It is used for measuring the similarity between two images. The difference with other techniques such as MSE or PSNR is that these approaches estimate absolute errors. Structural information is the idea that the pixels have strong inter-dependencies especially when they are spatially close. These dependencies carry important information about the structure of the objects in the visual scene.

The SSIM index is calculated on various windows of an image. The measure between two windows x and y of common size $N \times N$ is given by

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where μ_x, μ_y are the averages of the windows x and y, σ_{xy} is the covariance of x and y, $\sigma_x^2 and \sigma_y^2$ are the variances of x and y and c_1, c_2 are two variables to stabilize the division with weak denominator.

The resultant SSIM index is a decimal value between 0 and 1, and value 1 is only reachable in the case of two identical sets of data and therefore indicates perfect structural similarity. A value of 0 indicates no structural similarity.

5.2 Baseline Model

Before proceeding as per our proposed approach, we initially trained the CycleGAN [7] model without any changes in the Loss function as our baseline ($\lambda_c = 10$ is fixed throughout).

We train the model for 50 epochs using Adam optimizer and a learning rate of 0.0002. Also, as suggested in the CycleGAN paper, we update the discriminator using an image buffer of 50 previously generated images.

5.3 Tuning of $\lambda_{perceptual}$

Keeping the rest of the hyperparameters same as in the baseline model, we tune the weight of the perceptual loss (λ_p) to obtain the best possible results.

The following training specifications were used while tuning of the hyper-parameter λ_p:

1. To save the training time, we used only 200 blind image pairs and trained the model for 50 epochs for each value of $\lambda_p \in \{1, 5, 10\}$

2. After training, the model was evaluated on validation dataset for all values of λ_p.

Model	λ_p	**#Epochs trained**	**Average SSIM**
Proposed model	1	50	0.8545
Proposed model	5	50	**0.9825**
Proposed model	10	50	0.8692

Hence, the optimal value chosen for further analysis of λ_p is 5.

6 Comparison between the baseline model and proposed model

Finally, we train our proposed model with $\lambda_p = 5$ on the full dataset for 50 epochs. The comparison between the proposed model and the baseline on the test dataset is as follows:

Model	**Average SSIM**
Baseline	0.9894
Proposed	0.9827

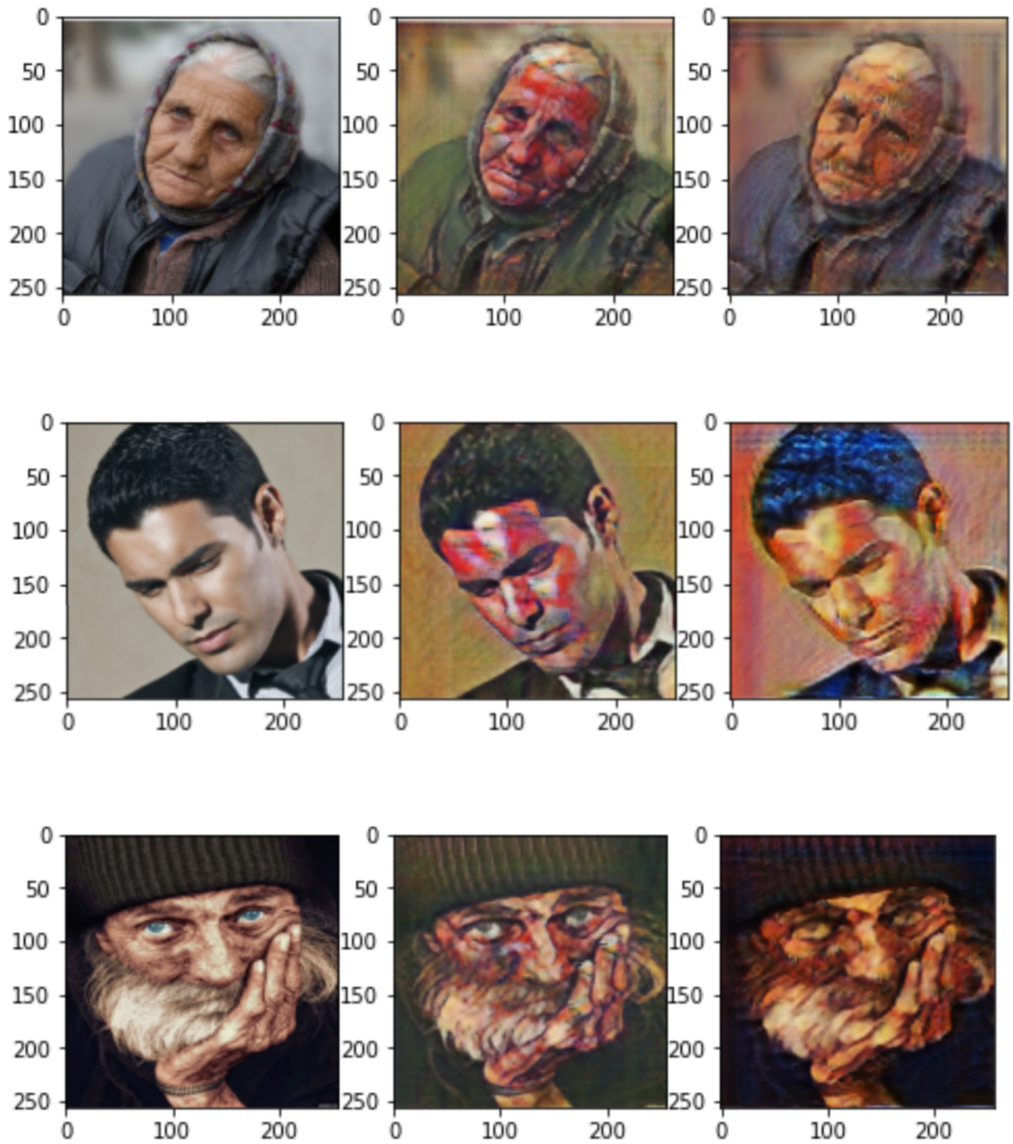


Figure 1: Image, Portrait from Baseline, Portrait from Proposed Model

7 Conclusion

In this project, we attempted to modify the CycleGAN model using facial features from facenet model to create better portraits from human photos. We observe that the original CycleGAN model still does slightly better that our proposed model in terms of the SSIM metric. Although our model was able to produce more visually appealing results in some cases, the overall performance of both models seems to be similar.

References

[1] Yuke Fang, Jiani Hu, and Weihong Deng. Identity-aware cyclegan for face photo-sketch synthesis and recognition. *CoRR*, abs/2103.16019, 2021.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

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

[6] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017.

[7] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. *CoRR*, abs/1609.03552, 2016.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.