Unpaired Image to Image Translation with Perceptual Style Loss

Gunjan Payal
2023H1030087P

Adwait Gondhalekar
2023H1030089P

Birla Institute of Technology and Science, Pilani

Abstract

The problem we chose was unpaired image to image translation i.e. style transfer without explicitly pairing images. This is done by using a CycleGAN architecture. We chose this architecture since it's domain independent and the generators for both the domains are trained at once.

1 Introduction

The goal in image to image translation is to learn a function $G : X \rightarrow Y$, such that the distribution of images from $G(X)$ is indistinguishable from the distribution of images in X. The way the reference paper [5] accomplished this was to simultaneously learn an inverse mapping $F : Y \rightarrow X$ and then by introducing a cyclic consistency loss.

The architecture of our reference paper is shown below:

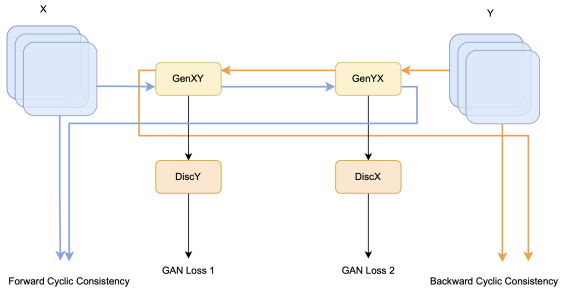

Figure 1: Reference Paper Architecture

The discriminator is a PatchGAN [2] discriminator with $70x70$ receptive field while the generator is a UNet [4] generator.

2 Loss Terms

The 3 terms in the loss are:

1. **Adversarial loss**: Adversarial loss is for the mapping $G : X \rightarrow Y$ is defined as:

$$\begin{aligned}&\mathcal{L}_{GAN}(G, D_Y, X, Y)\\&= \mathbb{E}_{y \backsim p_d(y)}[logD_Y(y)]\\&\quad + \mathbb{E}_{y \backsim p_d(x)}[log(1 - D_Y(G(x)))]\end{aligned} \tag{1}$$

2. **Cyclic consistency loss**: The cyclic consistency loss is defined as:

$$\begin{aligned}\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \backsim p_d(x)}[||F(G(x)) - x||]\\ + \mathbb{E}_{y \backsim p_d(y)}[||G(F(y)) - y||]\end{aligned} \tag{2}$$

3. **Identity loss**: The identity loss is defined as:

$$\begin{aligned}&\mathcal{L}_{idt}(G, F, x, y)\\&= \mathbb{E}_{y \backsim p_d(y)}[||G(y) - y||] + \mathbb{E}_{x \backsim p_d(x)}[||F(x) - x||]\end{aligned} \tag{3}$$

The final loss for the generator is then defined as:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X)\\ + \mathcal{L}_{cyc}(G, F) + \mathcal{L}_{idt}(G, F, x, y)\end{aligned} \tag{4}$$

3 Our Work

To adopt the original architecture for "artistic" style transfer, we introduced an additional loss term inspired from the original style transfer paper[1] as well as its modified version [3].

We compute the gram matrix (style encoding) from the target image and the gram matrix from the generated image and compute their L_1 difference and use that as another loss in the generator loss equation.

Let $K_l : X \rightarrow Z^l$ be a function that takes in an input image and produces the gram matrices of the image where l is the number of intermediate convolutional layers that the gram matrix is taken from. The loss can then be written as:

$$\mathcal{L}^{fw}_{style}(G, X, Y) = \mathbb{E}_{x \backsim p_d(X), y \backsim p_d(Y)}[||K(y) - K(G(x))||] \tag{5}$$

$$\mathcal{L}^{bw}_{style}(F, Y, X) = \mathbb{E}_{x \backsim p_d(X), y \backsim p_d(Y)}[||K(x) - K(F(y))||] \tag{6}$$

So, the total loss for the generator can be defined as:

$$\begin{aligned}\mathcal{L} = \lambda_{gan}[(\mathcal{L}_{GAN}(G, D_Y, X, Y) + \lambda_{style}\mathcal{L}^{fw}_{style}(G, X, Y)) \\ + (\mathcal{L}_{GAN}(F, D_X, X, Y) + \lambda_{style}\mathcal{L}^{bw}_{style}(F, Y, X))] \\ + \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{idt}(\mathcal{L}_{idt}(G, y) + \mathcal{L}_{idt}(F, x))\end{aligned} \tag{7}$$

Where, λ_{gan}, λ_{style}, λ_{cyc}, λ_{idt} are hyperparameters that serve as regularization terms in the loss.

The architecture is detailed in the figure below:

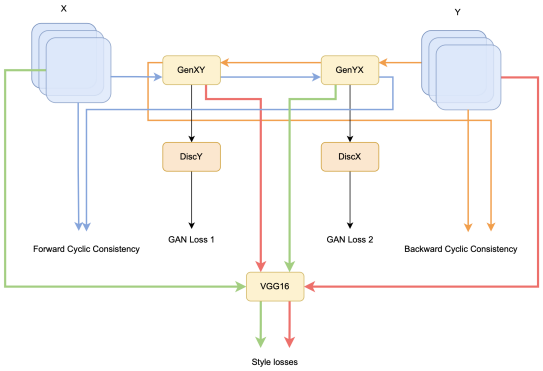

Figure 2: Our Architecture

4 Training Environment

The model was trained on an A100, 40GB VRAM.

The dataset was created by writing a simple script and crawling the WikiArt website.

The training set size was 634 cubism images and 1910 impressionistic images.

5 Results

Some results from running the model are shown in Fig. 3. Note that each pass produces style transferred images over both domains ("Transformed X", "Transformed Y") and the reconstruction done for the cyclic consistency loss i.e. $F(G(x))$ ("Reconstructed X") and $G(F(x))$ ("Reconstructed Y") are also shown.

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[3] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

6 Results (Images)

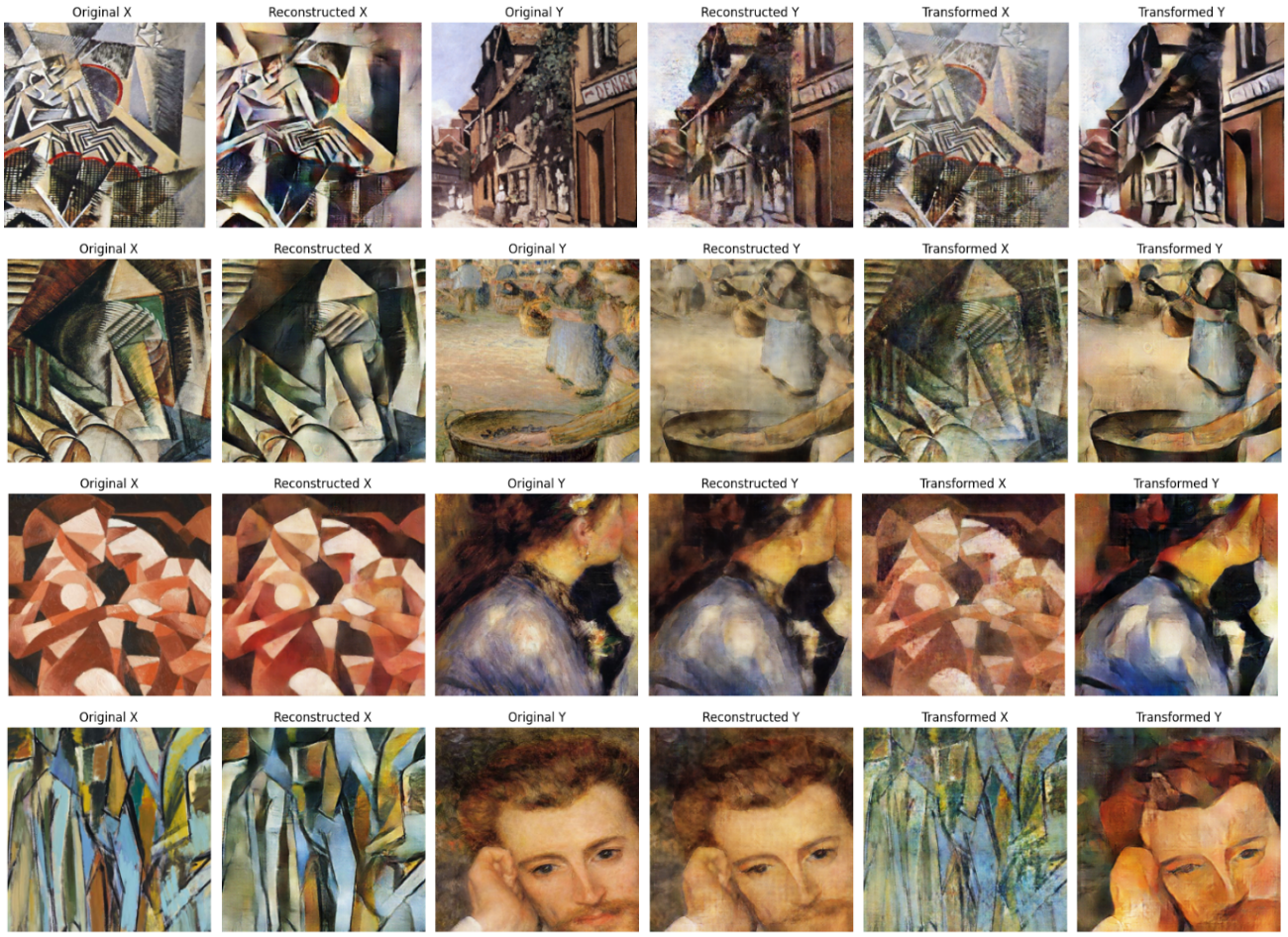

Figure 3: Generated Images