1081 Deep Learning - Final Project Report

108753130 資科碩一 李宣毅

107753005 資科碩二 湯昊軒

107753033 資科碩二 李柏彥

Implementation of Cycle-consistent Generative Adversarial Networks

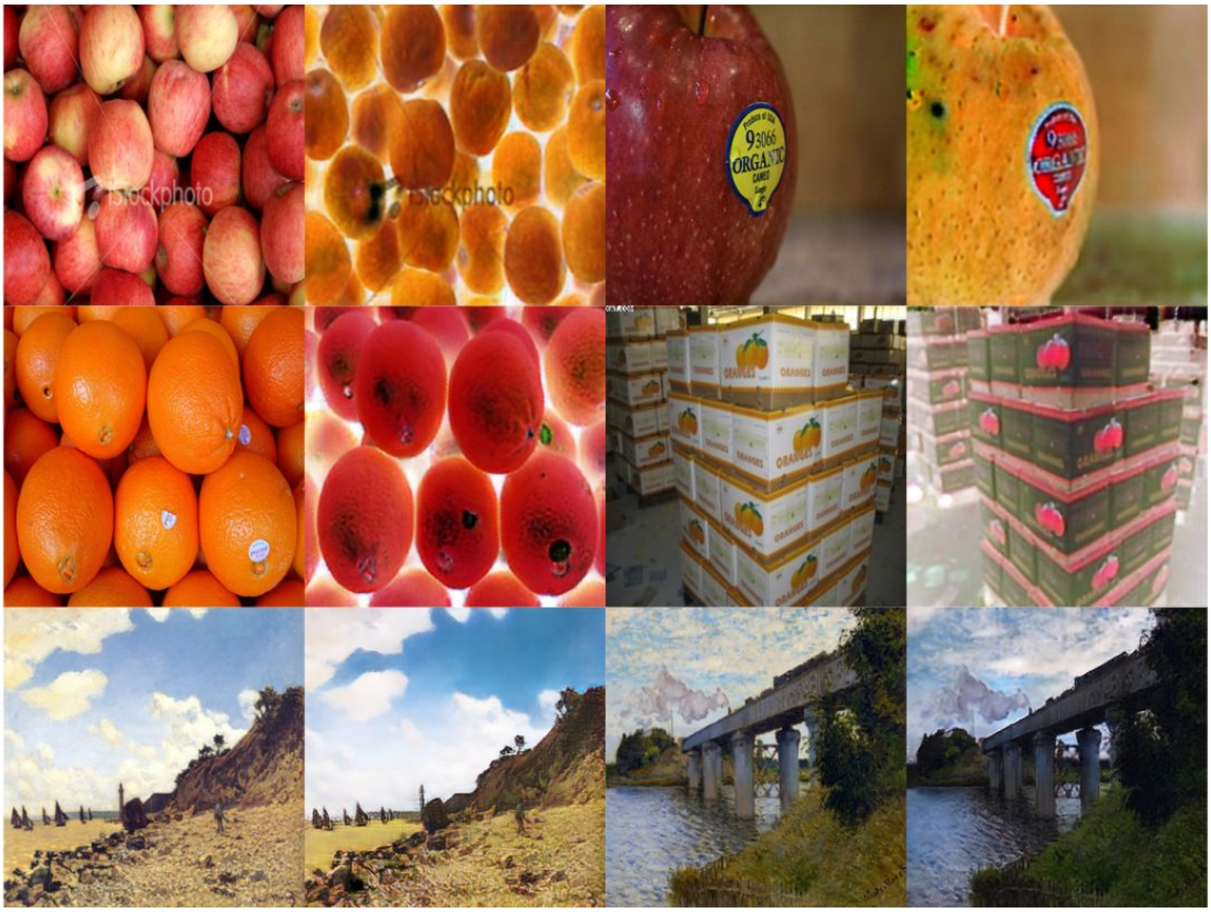

Figure 1: Representative Image. We train two datasets: "Oranges to Apples" and "Monet Paints to Photos" by out implementation of CycleGAN model. The first row is the translation of apples to oranges, the second row is the translation of oranges to apples, and the third row is the translation of Monet Paints to Photos. The image on the left side is the input image, and the right side is the output image.

Abstract

Since Generative Adversarial Networks(GAN) has been proposed in 2014, more and more computer vision tasks have been solved. There is even a GAN-Zoo online which records all kinds of GAN proposed in 2016~2017. Image-to-image translation is a class of vision and graphics

problems where the goal is to learn the mapping relations between an input image and an output image using a training set of image pairs. However, for many kinds of tasks, paired training data will not be accessible easily. So we implement a GAN model, CycleGAN, which is proposed in “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks” in 2017 ICCV conference. This model is for learning to translate an image from an input domain X to a target domain Y in the lack of paired training data.

1. Introduction

Image-to-image translation dates back to “Image Analogies” which is proposed in 2001, the research involves two stages: first, the system learns the mapping between a paired input image, and second the system filter another image by what it learns. After deep convolutional GAN(dcGAN) is proposed, unsupervised learning of GAN gets more attention and more researchers engage in the development of GAN.

With paired data in GAN training, it’s easy to find how an input image that belongs to domain X should be translated to another domain Y. With unpaired data in training, it still seems to find how an input domain should be translated, but something wrong may happen: no matter what you input to the model, it will output an image in training sets of target domain Y which is completely unrelated to your input, this phenomenon is called “Mode Collapse”. CycleGAN is proposed for the avoidance of mode collapse in the unsupervised learning. It uses the concept of “cycle

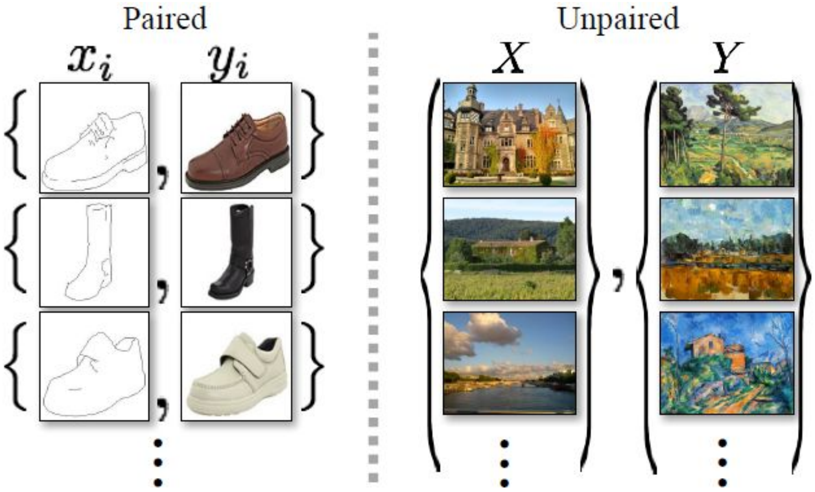

Figure 2: Paired training data (left) consists of training examples { $\{x_i, y_i\}_{i=1}^{N}$, where the correspondence between x_i and y_i exists. We instead consider unpaired training data (right), consisting of a source domain set $\{x_i\}_{i=1}^{N}$ ($x_i \varepsilon\, domain\, X$) and a target domain set $\{y_j\}_{j=1}^{M}$ ($y_j \varepsilon\, domain\, Y$), with no information provides as to which x_i matches which y_j.

consistency", in the sense that if we translate. e.g., a sentence from English to Chinese, and then translate it back from Chinese to English, we should arrive back at the original sentence. So in GAN model, if we have a translator G : domain X → domain Y ,we need to train another translator F : domain Y → domain X, which is the inverse of G. Then we should add a cycle consistency loss that encourages F(G(X)) ≈ X and G(F(Y)) ≈ Y. Combining this cycle-consistency loss with adversarial losses yields the objective for unpaired image-to-image translation.

2. Related Works

Image-to-image Translation Our approach builds on the "pix2pix" framework of Isola et al. [1] that can use same method to deal with a lot of things rather than redesign a loss function to map our goal. And uses a conditional generative adversarial network [2] to learn a mapping from input to output images.

Unpaired Image-to-Image Translation The goal is to relate two data domains: X and Y. Concurrent to our method, Liu et al. [3] offer a Unsupervised translation framework with a combination of variational autoencoders [4] and generative adversarial network [2] .

Cycle Consistency Verifying and improving translation via "back translation and reconciliation" is technique used by human translators [5], as well as machines [6]. Concurrent with our work, Yi et al. [7] independently use a similar objective for unpaired image-to-image translation, inspired by dual learning in machine translation [6].

Neural Style Transfer [8] is another way to perform image-to-image translation which synthesizes a novel image by combining the content of one image with the style of another image used by convolutional neural networks.

3. Methods

3.1 Formulation

Our goal is to learn mapping functions between two domains X and Y given training samples $\{x_i\}_{i=1}^{N}$ $(x_i \varepsilon X)$ and $\{y_j\}_{j=1}^{M}$ $(y_j \varepsilon Y)$. As illustrated in Figure 3 (a), our model includes two mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$. In addition of GAN model, we introduce two adversarial discriminators D_X and D_Y , where D_X aims to distinguish between images $\{x\}$ and translated images $\{F(y)\}$; in the same way, D_Y aims to discriminate between $\{y\}$ and $\{G(x)\}$,Our objective contains two types of terms: adversarial losses for matching the distribution of generated images to the data distribution in the target domain; and cycle consistency losses to prevent the learned mappings G and F from contradicting each other.

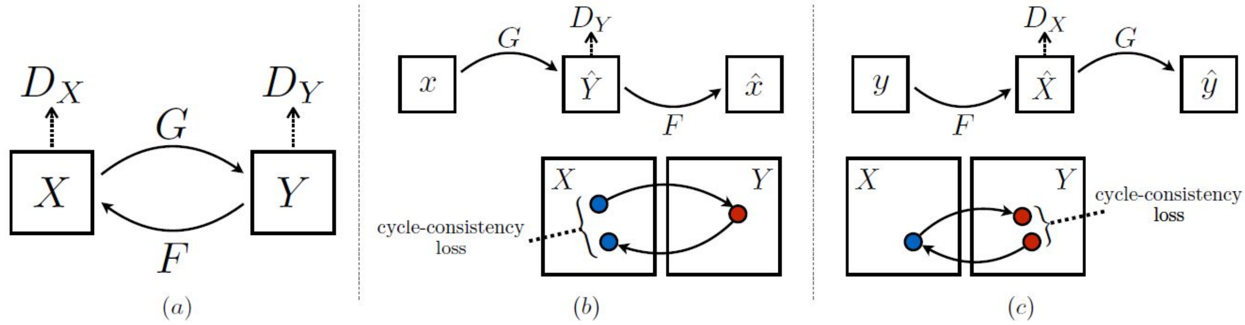

Figure 3: (a) our model includes two mappings functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_X and D_Y

.(b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

3.1.1 Adversarial Loss

We apply adversarial losses to both mapping functions. For the mapping function $G : X \rightarrow Y$ and its discriminator D_Y , we express the objective as:

$$L_{GAN}(G, D_Y, X, Y) = E[log\ D_Y(y)] + E[log(1 - D_Y(G(x))] \qquad (1)$$

where G tries to generate images $G(x)$ that look similar to images from domain Y , while D_Y aims to distinguish real or not between translated samples $G(x)$ and real samples y . G aims to minimize this objective against an adversary D that tries to maximize it.

3.1.2 Cycle Consistency Loss

Adversarial training can, in theory, learn mappings G and F that produce outputs identically distributed as target domains Y and X respectively. If data are paired, it’s easy to find how the special characteristics of the input should be translate to the labeled output. If data are not paired, it seems to generate y’ which is indistinguishable from domain Y after G.

However, the G network may not be what you think. Only with the adversarial training in the unpaired data training, a generator can map the input images to any random of images in the target domain, and the discriminator will always consider it to be *real*. This phenomenon is called Mode Collapse.

To avoid this condition, the mapping function should be cycle-consistent: as shown in Figure 3 (b) and (c), the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ and $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ .So the model should be trained using a cycle consistency loss:

$$L_{cyc}(G, F) = E[\|F(G(x)) - x\|_1] + E[\|G(F(y)) - y\|_1 \qquad (2),$$

which the loss is used the L1 norm.

3.1.3 Objective Function

the objective function is:

$$L = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_x, Y, X) + \alpha L_{cyc}(G, F) \qquad (3)$$

where α controls the relative importance of the two objectives.

3.2 Implementation

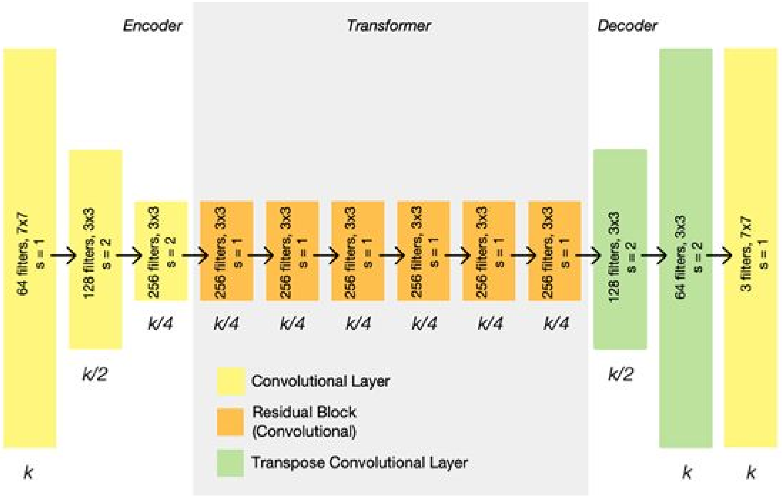

Figure 4: generator network architecture.

Generator Network Architecture The architecture of the generator networks is U-net, as shown in Figure 4. This network contains two stride-2 convolutions, 9 redisual blocks, and two strided $\frac{1}{2}$ convolutions. we use instance normalization instead of batch normalization because instance normalization is proved to get the better result than batch normalization in the GAN research which aim to generate natural image. The input shape of the network is set to 256×256. To sum up, the first half of the U-net is the feature extraction, and the second half is up-sampling. And in the network, we have about 2.8 million of trainable parameters.

Discriminator Network Architecture The architecture of the discriminator networks is the discriminator of PatchGAN, which is a simple convolutional network. The difference of D between the PatchGAN and the normal GAN model is, the output of the D in the normal GAN is 1 (real) or 0 (not), and the output in the PatchGAN is a 2-dimensional array of $1s$ or $0s$, which means every region (called patch) is real or not. The advantages of the PatchGAN is in the condition that when any patch's loss are larger than others while training, the model can update the parameters relative to the patch instead of update all the parameters. it gets a better performance and a faster convergence of the loss function.

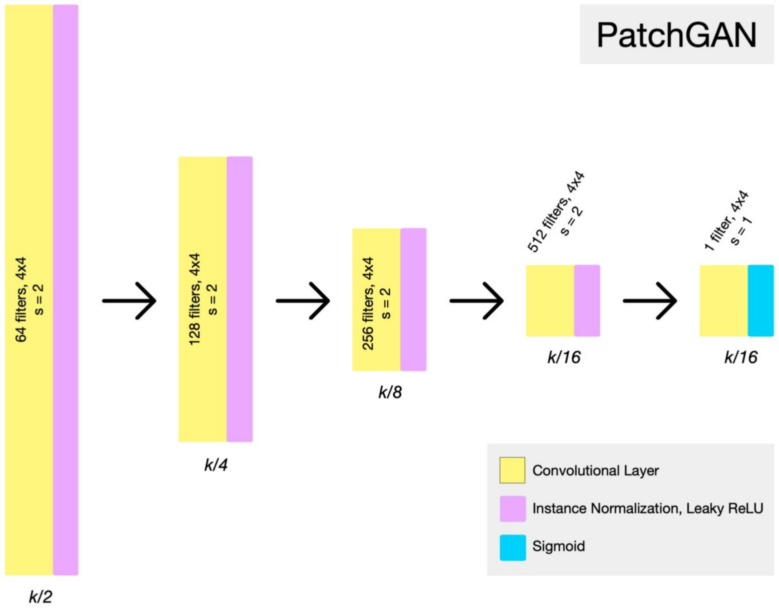

Figure 5: discriminator network architecture.

Training Details In the Consideration of time, we only trains 100 epochs with fixed learning rate $lr = 0.0002$ instead of 200 epochs (as the paper mentioned), and we only train the Apples ↔ Oranges and Monet ↔ Photos datasets.The numbers of image in the apples and oranges are about to one thousand, and that in the Monet paints and real photos are about to seven hundreds and six thousands.

4. Results

The translation results is shown below (Figure 6, 7, 9). We set α in the formula (3) to 10 when training. All the left side is the input image and the right side is the output image.in the Apples↔Oranges datasets, we trained the model for about twelve hours. In the training of Monet↔ Photos, we limited the iteration (2000 iterations per epoch) to reduce the training time, and it costs about eighteen hours, if we don't use the limitation, it will costs more than sixty hours.

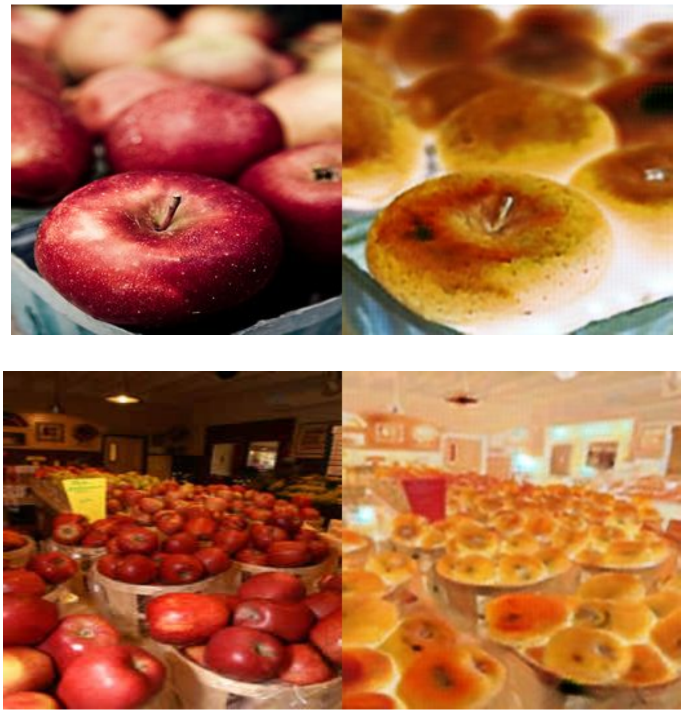

Figure 6: the translation of Apples to Oranges.

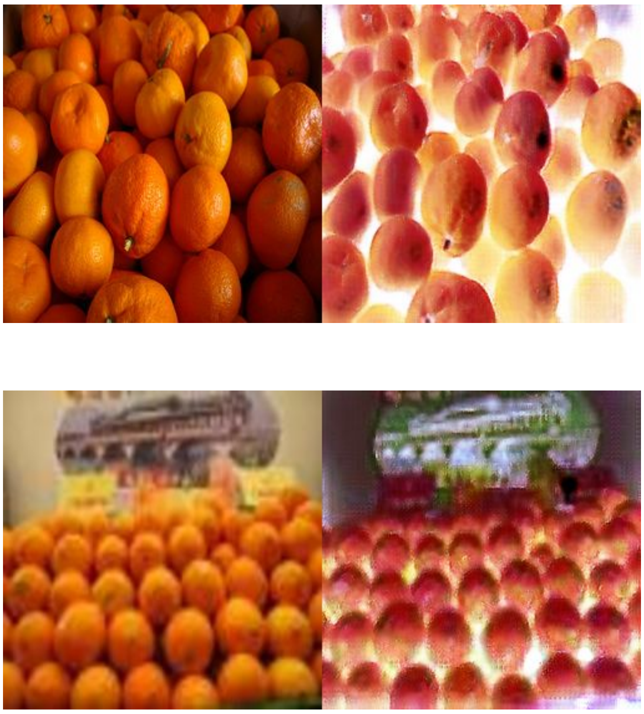

Figure 7: the translation of Apples to Oranges.

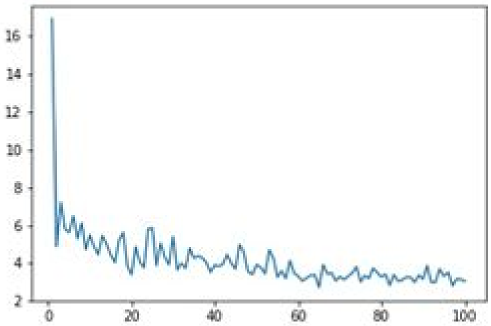

Figure 8: the loss diagram in the training of Apples ↔ Oranges datasets.

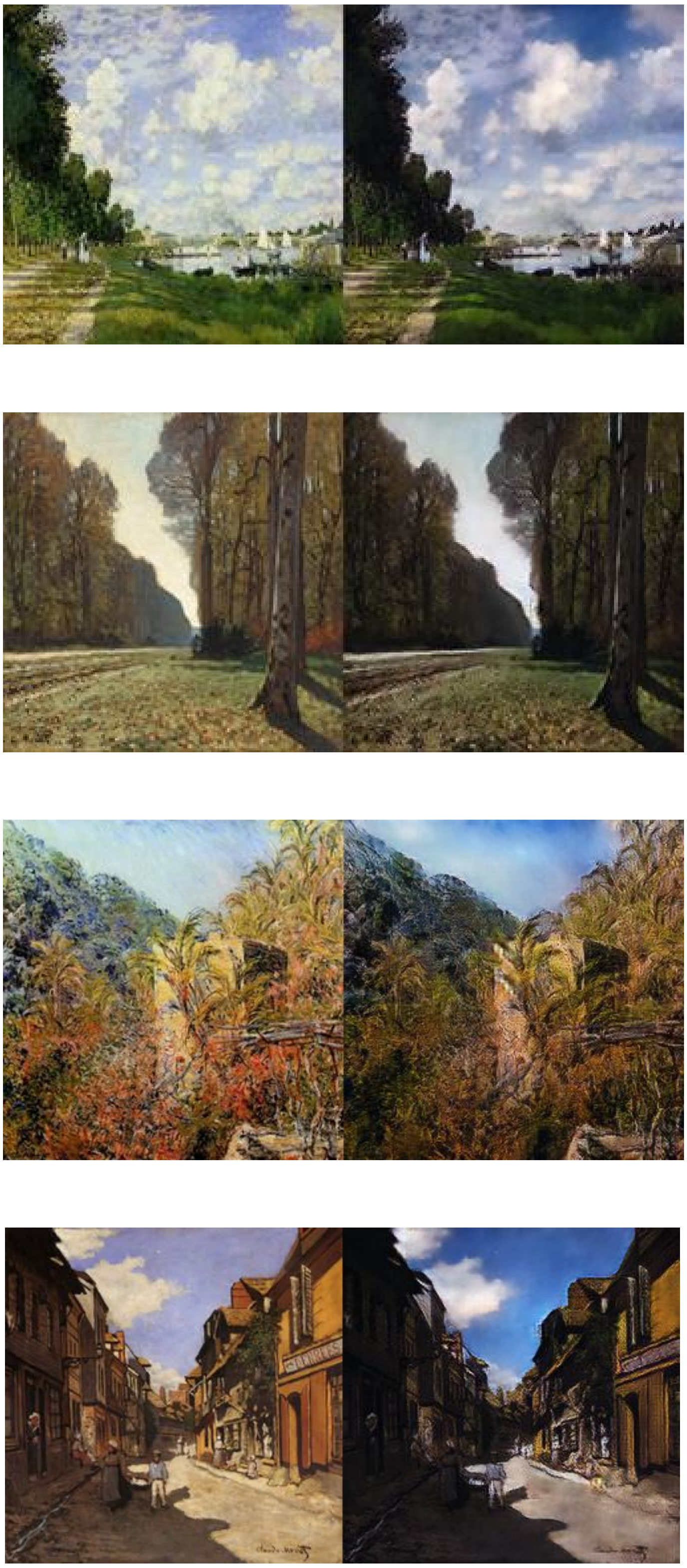

Figure 9: the translation of Monet paints to real photos.

Identity Learning With the proposed GAN architecture, it can avoid the condition of Mode Collapse. However, the pixel value in the background or non-region-of-interest region (non-ROI) of the output may be totally different. It's because that the color of the background in the training set of domain X is different from that of domain Y . In other words, if the non-ROI of the Apples set is dark and that of the oranges set is bright, it will generate an output like Figure 10.

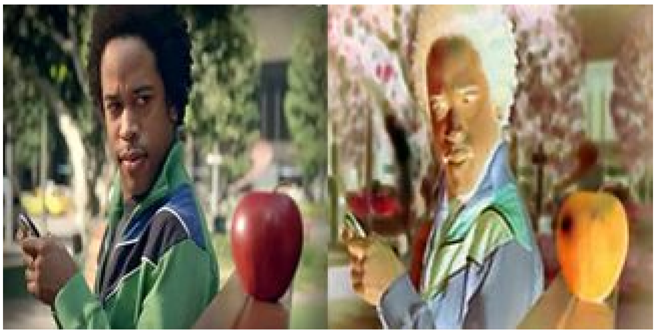

Figure 10: an output during the apple to orange translation. The background (non_ROI) of the input image get the opposite output.

So we add additional identity loss $L_{identity}$:

$$L_{identity}(G, F) \ = \ E[||G(y) - y||_{\,1}] \ + \ E[||F(x) - x||_{\,1} \qquad (4)$$

to the objective loss function of the CycleGAN model. The results is shown in Figure 11 ~ 12.

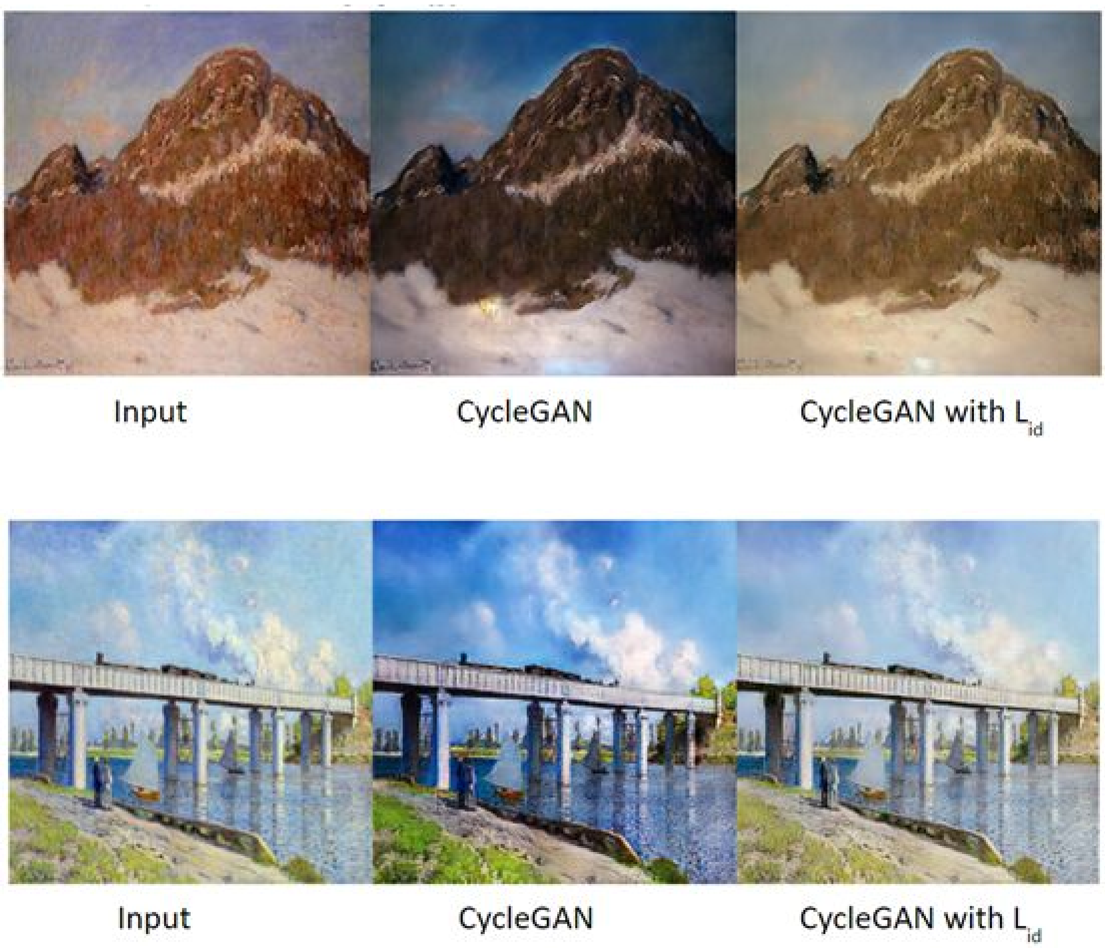

Figure 11: the comparison of the generated outputs with or w/o identity loss. The images are “Monet” inputs, the “Photo” outputs generated by normal CycleGAN model, and the “Photo” outputs generated by the CycleGAN with identity learning model from the left side to the right side. The pixel value in the non-ROI region is more similar to original input.

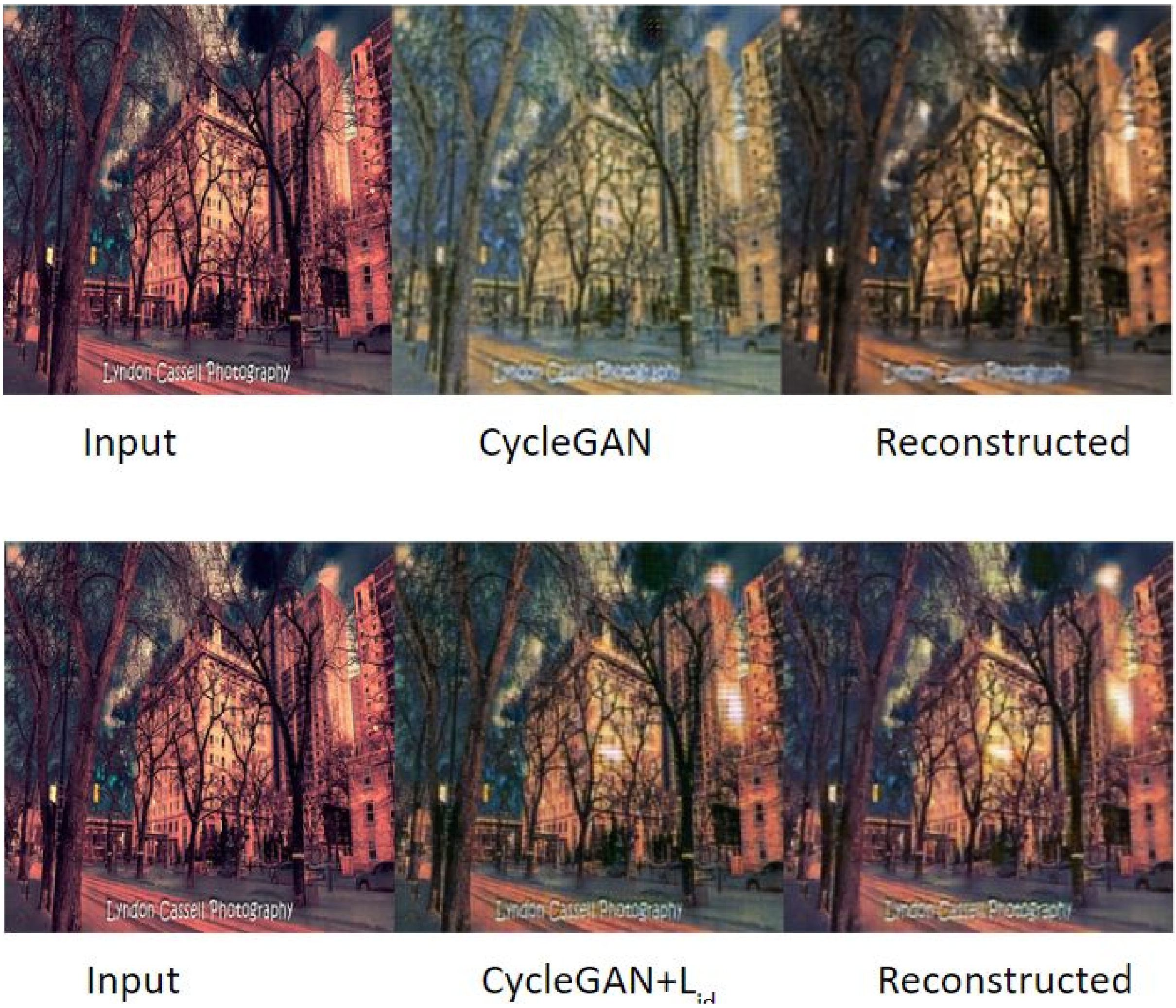

Figure 12: the comparison of the reconstructed outputs with or w/o identity loss. The images are “Photo” inputs, the “Monet” outputs reconstructed by normal CycleGAN model, and the “Monet” outputs reconstructed by the CycleGAN with identity learning model from the left side to the right side.

5. Discussion

In the Consideration of time, we only trains 100 epochs with fixed learning rate $lr = 0.0002$, though the paper mentioned that they train 200 epochs with fixed learning rate at the first half and linearly descended learning rate at the second half. Figure 8 show the loss diagram of the training, and it reveals that only 100 epochs may not reach to the real converge of the objective loss function. And the outputs in the results section are far from that of the original paper.

And the outputs in the results section is the better outputs than other images. more images in the outputs is far from great results.

After the addition of identity loss (mentioned in results section), the model solve the non-ROI regions' problem, but the output of the model extremely looks the same with the original input. this derived problem implies that maybe we implement a wrong way of the addition of identity loss, and the paper doesn't detailed how they use this loss function to train and optimize the model.

6. Conclusion

This paper proposed an learning model in the lack of the paired training data, which called "CycleGAN". It use the concept of "Cycle-consistency" to build up the loss function and the model, which makes the output avoid from "Mode Collapse".

Limitation CycleGAN can only make minimal changes to the input, and the geometry of the training sets should be similar (apples and oranges are like sphere), if we want to train a translation between dogs and cats, it will gets failure (maybe a dog facial features on a cat's face...etc).

Nonetheless, in many cases completely unpaired data is plentifully available and should be made use of. This paper pushes the boundaries of what is possible in this “unsupervised” setting.

7. Reference

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

[3] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In NIPS, 2017.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. ICLR, 2014.

[5] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W.-Y. Ma. Dual learning for machine translation. In NIPS, 2016.

[6] R. W. Brislin. Back-translation for cross-cultural research. Journal of cross-cultural psychology, 1(3):185–216, 1970.

[7] Z. Yi, H. Zhang, T. Gong, Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In ICCV, 2017.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. CVPR, 2016.

[9] Reflection Padding. Retrieved from https://stackoverflow.com/questions/50677544/reflection-padding-conv2d