Optimization of CycleGAN Model: Final Report

G071 (s2056595, s2123157, s2196789)

Abstract

The purpose of image-to-image translation is to learn the mapping between an input picture and an output image using a training set of matched image pairings. When matched training data is unavailable, CycleGAN provides a method for learning how an image may be translated from one source domain to another. While CycleGAN's strategy often produces persuasive findings, the outcomes are far from universally good. Certain types of failures are triggered by the distribution properties of the training datasets. Additionally, we notice a persistent discrepancy between the outcomes obtained using paired training data and those from unpaired strategy. Managing increasingly complex and intense transformations, particularly geometrical transformations, is the main purpose of our evaluation of a new modified unsupervised generative model with attention. Unpaired data is not necessary for unsupervised generative networks, which is reflected in the CycleGAN structure. Weights in the feature map indicate attention, which can partition the source domain from the target domain using the auxiliary classifier's attention map. We present an adaptive hybrid normalisation layer composed of LN and IN to assist attention-guided models in controlling the amount of shape and texture variation without adjusting the model architecture or hyper-parameters.

1. Introduction

While the image to image translation problems has been widely investigated since the first generation of Generative Adversarial Networks(GAN) was invented in June, 2014 by Ian Goodfellow et al, the most state-of-art method for this field of tasks is still being pursued by people year by year(I. Goodfellow, 2014a). Image-to-image translation has a wide range of applications including changing a carton photo to a realistic photo, or converting real landscape map to satellite map. However, at the early stage of the work, people were required to collect large amount of paired dataset for training neural networks, which is called pix2pix(P. Isola, 2016). In reality, it is usually very difficult to collect paired images due to the constraints of different situations, and hence the generated results are not quite satisfactory. Purposed by Jun-Yan Zhu et al, in 2017, CycleGAN had greatly improve the performance in tasks of image-to-image translation, and became very popular as a state-of-art method(J. Zhu, 2017).

The most significant improvement in CycleGAN is that it doesn't require paired images, which means we only need to care about one kind of images if we want to translate from A to B.The architecture of CycleGAN consists of two Generative models and two discriminator models. The function of generator G maps from X to Y such that the distribution of X are reproduced in the generated results $G(X)$, and the generator F works exactly in the opposite direction.

There are some limitations about CycleGAN. Although it has good performance on tasks involving texture or color transformation, such as converting horses to zebras, it does not perform well on geometrical translations. In some examples of geometrical translations it fails to translate from cat to dog. Some distributions of features would also lead to failure, as CycleGAN is incapable of learning multimaps from unpaired data (A. Almahairi, 2018). For instance, when translating from horses to zebras, it doesn't consider there could be a man on the horses and as a result the man was also covered with zebra stripes(Saxena, 2021).

In this report we aim to assess a new modified unsupervised generative model with attention. We will test, verify, and evaluate the enhanced metrics and optimization effects in comparison to CycleGAN based on the model's principle of Generative Adversarial Networks.

This study's findings and indicators are excellent. The following are the primary features of this modified model with attention that are worth noting:

• Unsupervised Generative Networks: Reflected in the CycleGAN structure, unpaired data is not required.

• Attention (K. Xu, 2015): It is reflected in the feature map with weight. The specific tactic is to assist the model in determining where to focus the transformation by separating the source domain from the target domain through using attention map obtained by auxiliary classifier.

• Adaptive Normalization Layer: An adaptive hybrid normalization layer of LN and IN is introduced to help our attention-guided model flexibly control the amount of shape and texture variation without modifying the model architecture or hyper-parameters.

2. Data set and task

In this report, we use multiple datasets to accomplish different image-to-image translation tasks based on our modified model. We use the same data as used in CycleGAN (EEC, 2021) **horse2zebra**, **apple2orange**, **summer2winter**, and **iphone2dslr_flower** . Additionally, we introduce extrinsic dataset like **cat2dog** (Asirra, 2017). We also use **selfie2anime** for testing the style transfer. For data set in each task, we separate training set and test set for two domains in our task. For each domain, we choose around 1000 images as training set and around 100 images as test set.

We aim to explore three representative image-to-image translation tasks which separately focusing on geometric changes, texture and color changes, and style changes. For each task, We specifically choose appropriate datasets to achieve it goals.

In task of geometric changes, we use **cat2dog** and **apple2orange** dataset to convert cats to dogs and convert apple to orange. **cat2dog** and **apple2orange** are datasets used to evaluate CycleGAN's proposed future optimization direction, which may successfully assess the function of attention in the conversion process via our modified model.

In task of texture and color changes, for **horse2zebra** datasetwe convert horses to zebra by changing textures and color of horses in the photo. For **summer2winter** and **iphone2dslr_flower** dataset to change.

In task of style changes, we use **selfie2anime** to convert photo styles from selfie to anime. The anime set includes several figures of anime characters with different drawing styles.

3. Methodology

3.1. CycleGAN

Our objective is to discover mapping functions between two domains X and Y using training data $\{x_i\}_{i=1}^N$ where $x_i \in X$ and $\{y_i\}_{j=1}^N$ where $y_i \in Y$. The data distribution is denoted by the variables $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. Our model, as indicated, has two mappings: $G : X \rightarrow Y$ and $F : Y \rightarrow X$. Additionally, we present two adversarial discriminators D_X and D_Y, where D_X seeks to differentiate between pictures $\{x\}$ and translated images $\{F(y)\}$ and D_Y seeks to distinguish between images $\{y\}$ and $\{G(x)\}$. Our aim includes two kinds of terms: *adversarial losses* for matching the distribution of produced pictures to the distribution of data in the target domain; and *cycle consistency losses* for preventing the learned mappings G and F from contradicting one another.

3.1.1. Adversarial Loss

Both mapping functions are susceptible to adversarial losses(I. Goodfellow, 2014b). The aim for the mapping function $G : X \rightarrow Y$ and its discriminator D_Y is as follows:

$$\mathcal{L}_{GAN}(G, D_X, Y, X) = \mathbb{E}_{y \sim p_{data}(y)}[log D_Y(y)]$$

$$+\mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))]$$

in which G attempts to produce pictures $G(x)$ that resemble images from domain Y, and D_Y attempts to differentiate between translated samples $G(x)$ and genuine samples y. G seeks to reduce this goal whereas D seeks to enhance it, i.e., $min_G\ max_{D_Y}\ \mathcal{L}_{GAN}(G, D_Y, X, Y)$. We also include an adversarial loss for the mapping function $F : Y \rightarrow X$ and its discriminator D_X: $min_F\ max_{D_X}\ \mathcal{L}_{GAN}(F, D_X, Y, X)$.

3.1.2. Cycle Consistency Loss

It is theoretically possible to learn mappings G and F that yield identically distributed outputs as target domain Y and X, respectively, using adversarial training (strictly speaking, this requires G and F to be stochastic functions)(Goodfellow, 2016). Any random permutation of pictures in the target domain, in contrast, may induce an output distribution that matches the target distribution when a network's capacity is high enough. A learnt function's ability to map an individual input x_i to a desired output y_i cannot be guaranteed by adversarial losses alone. We suggest that the learnt mapping functions should be cycle-consistent in order to further minimise the number of viable mapping functions. We use a loss in cycle consistency as a reward for this conduct:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1]$$

$$+\mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]$$

Reconstructed pictures $F(G(x))$ end up matching closely to the input images x because of cycle consistency loss.

3.1.3. Full Objective

Ultimately, our goal is to accomplish the following:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y)$$

$$+\mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F)$$

In this case, λ is used to regulate the relative relevance of the two goals. We're here to address:

$$G^*, F^* = arg\ min(G, F)\ max(D_X, D_Y)\ \mathcal{L}(G, F, D_X, D_Y)$$

Not that our model may be considered as training two "autoencoders"(G. E. Hinton, 2006): we learn one $F \circ G : X \rightarrow X$ and another $G \circ F : Y \rightarrow Y$ concurrently. Each of these autoencoders, however, has a unique internal structure: they map a picture to itself through an intermediate representation, which is a translation of the image into another domain. This configuration may alternatively be thought of as a subset of "adversarial autoencoders"(A. Makhzani, 2016), which use an adversarial loss to train an autoencoder's bottleneck layer to match an arbitrary target distribution. In our scenario, the $X \rightarrow X$ autoencoder's target distribution is the domain Y.

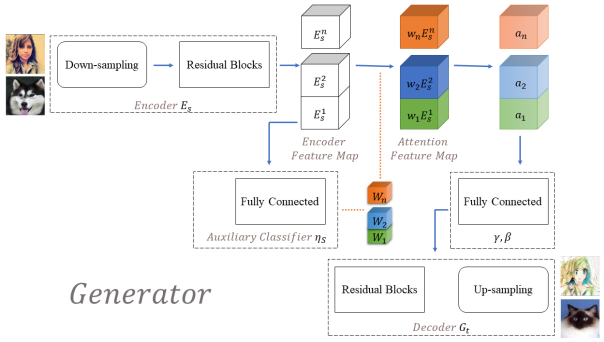


Figure 1. The structure of the **Generator**

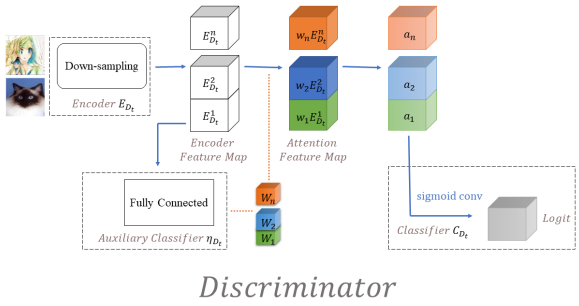


Figure 2. The structure of the **Discriminator**

3.2. Modified Unsupervised Generative Model with Attention

3.2.1. Generator

First, the image goes through a down-sampling module, and then through residual blocks to obtain the encoder feature map, which is divided into two paths. One is to obtain the weight information of each feature map through an auxiliary classifier, and then multiply it with the encoder feature map of the other path to obtain the attention feature map. The attention feature map is still divided into two paths. One goes through a 1x1 convolution and activation function layer to get the $a_1 ... a_n$ feature map, and then the feature map passes through the fully connected layer to get the γ and β of the Adaptive Layer-Instance Normalization layer in the decoder. The other path serves as the decoder's input, and the generated result is obtained through adaptive residual blocks (including Adaptive Normalization Layer) and an up-sampling module.

3.2.2. Adaptive Normalization Layer

Informed by recent research that employs affine transformation parameters in normalisation layers and the combination of normalisation functions (X. Huang, 2017)(H. Nam, 2018), the precise formula we provide is as follows:

$$\hat{a}_I = \frac{a - \mu_I}{\sqrt{\sigma_I^2 + \epsilon}}, \hat{a}_L = \frac{a - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}}$$

The above is the normalisation formula for IN and LN, after which $\hat{a}_I$ and $\hat{a}_L$ are substituted for merging (γ and β are passed in externally):

$$\text{AdaLIN}(a, \gamma, \beta) = \gamma \cdot (\rho \cdot \hat{a}_I + (1 - \rho) \cdot \hat{a}_L) + \beta$$

To avoid ρ from exceeding the range of [0, 1], the interval of ρ is trimmed as follows:

$$\rho \leftarrow clip[0,1](\rho - \tau \Delta \rho)$$

While AdaIN is capable of transferring content features to style features, AdaIN relies on the assumption that feature channels are uncorrelated (X. Huang, 2017), which indicates that style features should include a huge number of content patterns, whereas LN does not (J. L. Ba, 2016). However, because LN considers global statistics, it would be unable to maintain the content structure of the original domain. As a result, we integrate AdaIN and LN and leverage their combined capabilities to selectively preserve or update content information, therefore assisting in the resolution of a broad variety of image-to-image translation difficulties.

3.2.3. Discriminator

The discriminator is significantly simpler than the generator whose particular structure is similar to the generator, but it employs spectral normalisation to improve training stability and convergence, and it uses leakyrelu for the activation function.

3.2.4. Adversarial Loss

The adversarial loss function is not the original log function (X. Mao, 2017), but rather the MSE function.

$$L_{GAN}^{s \to t} = \left(\mathbb{E}_{x \sim X_t} \left[(D_t(x))^2 \right] + E_{x \sim X_s} \left[(1 - D_t(G_{s \to t}(x)))^2 \right] \right)$$

3.2.5. Cycle Loss

Under the CycleGAN architecture, the cycle consistency loss is as follows: A is translated to B, and then B is translated to A'; A and A' must be identical, and the loss is L1loss.

$$L_{Cycle}^{s \to t} = \mathbb{E}_{x \sim X_s} \left[|x - G_{t \to s}(G_{s \to t}(x))|_1 \right]$$

3.2.6. Identity Loss

Identity loss assures that the colour distributions of the input and output images are similar.

$$L_{Identity}^{s \to t} = \mathbb{E}_{x \sim X_t} \left[|x - G_{s \to t}(x)|_1 \right]$$

3.2.7. Class Activation Map (CAM) Loss

The purpose of CAM (J. Kim, 2019) is to utilise the information from the auxiliary classifiers η_s and η_{D_t} to determine where they need to be improved, or what the largest disparity between the two domains is in their current state, which $G_{s \to t}$ and D_t find out, given an image $x \in (X_s, X_t)$.

Model name	G_A(M)	G_B(M)	D_A(M)	D_B(M)
CycleGAN	11.378	11.378	2.765	2.765
Modified Model	10.59	10.59	53.13	53.13

Table 1. Number of trianable parameters of CycleGAN and our modified model.

The generator's CAM loss, calculated using BCE loss:

$$L_{\text{CAM}}^{s\to t} = -\mathbb{E}_{x\sim X_s}\left[\log\left(\eta_s(x)\right)\right] + \mathbb{E}_{x\sim X_t}\left[\log\left(1-\eta_s(x)\right)\right]$$

The discriminator model had a significant error in the formulation provided by the original paper we referenced, which we addressed and corrected. The discriminator's CAM loss as measured by MSE:

$$L_{cam}^{D_t} = \mathbb{E}_{x\sim X_t}\left[\left(\eta_{D_t}(x)\right)^2\right] + \mathbb{E}_{x\sim X_s}\left[\left(1-\eta_{D_t}\left(G_{s\to t}(x)\right)\right)^2\right]$$

4. Experiments

4.1. Overview

In this experiment we trained the models on two frameworks, the one with typical CycleGAN architecture and the one with some modification which is built upon the architecture of CycleGAN. In order to evaluate the performance of both frameworks, we ran experiments on both framework on the same dataset for comparison. Different dataset will be used for testing different scenarios. The dataset we used in this experiment are **horse2zebra**, **apple2orange**, **cat2dog**, **summer2winter**, **iphone2dslr_flower** and **selfie2anime**.

For the sake of controlling variables, we keep the hyperparameters the same for all dataset we used. The crop size was capped at 256 for each image in order to assure that the desire target is included in the image. In the training process we use minibatch with size of 10 at each iteration. We also tried to decrease the size of each image to increase the training speed, but the loss of generators was fluctuated and did not converge, so we choose 256 by 256 as the input dimension of our model. We use adaptive learning rate with initial value of 0.0001. The total number of trianable parameters are showed in table1.

4.2. Model Modification

There are some notable modifications that are made on typical CycleGAN architecture. Table1 shows that the number of parameters of discriminator in modified model is far greater than that of parameters of discriminator in CycleGAN. This is because we add two extra discriminators for multi-scale model, one for global and the other one for local. The discriminator for global has receptive field of 190 and 7 convolutional layers, and that for local has receptive field of 46 and 5 convolutional layers. The active layer after each convolutional layer is leaky ReLU. The reason why we use this multi-scale model is that we want to discriminate pictures in different scale.

We use the model template provided in the original CycleGAN source code to implement our custom model. Based on the original framework, we add the Channel Attention Module(CAM) which is adapted from UGATIT to both generator and discriminator(J. Kim, 2019). Using the CAM module, we adjust the loss function for generator and discriminator by linearly combining the loss from the CAM module and loss from the original model. Furthermore, we use the adaptive layer-instance normalization which helps our attentional model to flexibly control the changes of shape and texture of images without changing model architecture or hyperparameters.

Loss item	Weight
GAN	1
Reconstruction	10
Cycle	10
CAM	100
Other Hyperparameters	**Value**
Learning Rate	0.0001
Weight Decay	0.0001
Resblock Number	4
Discriminator Layer	6
Image Size	256×256

Table 2. Weights for different loss item.

4.3. Loss and other hyperparameter settings

Figure3 shows several attributes of the loss. This figure is produced by running 200 epoch on the **cat2dog** dataset. Since the loss of translating from A to B (represents as A) has similar feature with that of translating from B to A (represents as B), this figure only includes the loss for A. Except for **cam_A**, other four parts of loss is the same as those in CycleGAN. According to this figure, both **cam_A** and **G_A** are decreasing as training proceed, while others are keeping fluctuate. Particularly, **G_A** is calculated as the sum of loss of generator A itself, reconstruction loss, CAM loss and identity loss, after times their weights respectively.

Table2 shows the weights for every item of loss, as well as other hyperparameter settings. The GAN loss includes loss of generator and discriminator. The former is expected to generate results such that the output of the later is one(real). The reconstruction loss aims to reduce the difference of real image and reconstructed image because we expect the reconstructed image is consistent with the original one. We also use identity loss because we want the generator to make no changes to the target domain image. The CAM loss is also needed to be constrained in order to make the attention module work.

We use adaptive learning rate with initial value of 0.0001 and weight decay value of 0.0001 to inhibit the increase of parameter value in each iteration. Furthermore, we reduce the number of residual blocks from 9 to 4 in generator to accelerate the training process. On the contrary, we have 6 convolutional layers in discriminator comparing with 3 in that of CycleGAN model.

Figure 3. CAM loss, Identity loss, Cycle loss, Generator loss, and Discriminator loss of A, on **cat2dog** dataset

4.4. Experiment Results

4.4.1. Dataset: cat2dog

We firstly ran experiment on the **cat2dog** dataset to test the model's ability to deal with image translation tasks when involving shape changes. The model is trained for 200 epoch with 100 iterations per epoch. In each iteration, 10 images is randomly chose from the dataset and loaded into image pool. The image pool is used to provide data for model to generate fake image and perform gradient descent according to the loss function.

Figure5 is the result we got after 200 epoch. There are 7 rows in figure5 with each row representing the output of a batch of image. Entries in even rows is the heat map of those in odd rows generated by attention module. From top to bottom, entries in odd rows are real_A, idt_A, fake_B, rec_A respectively. To better illustrate this, considering a task translating from A to B, real_A represents the input sample A given to the generator G and then generates the fake image $G(A)$ which is fake_B. Real_B and fake_A is for the other generator F working on the other way around, which is translating from B to A. The rec_A represents $F(G(A))$ which aims to reconstruct A in order to evaluate the cycle loss of the whole system. It is essential to keep the cycle consistency which ensures we can reconstruct A using generator F. The idt_A and idt_B is identity mapping which represents $G(B)$ and $F(A)$ to ensure $G(B) = B$ and $F(A) = A$.

By looking at the generated results, we could conclude that our model successfully fulfilled cat to dog translation task despite there are some translated dog images that are not perfect and have some flows. We got very good results on identity mapping since the first row(real_ A) is almost the same as the third row(idt_A), which is generated using generator B to A with input A. The fifth row is fake_B. In a few samples the translation is good and it is difficult to tell whether the image is a real dog or not. However on some

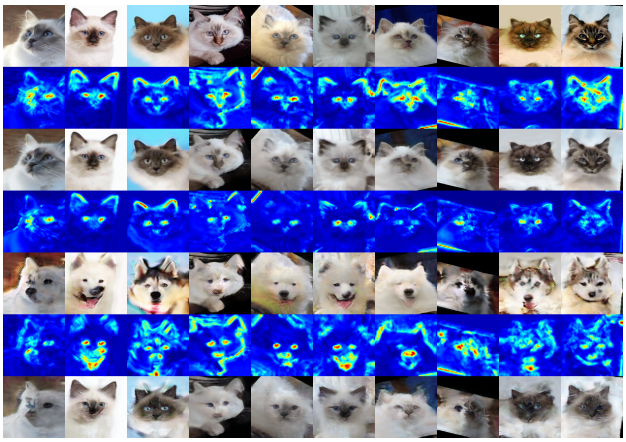


Figure 4. Generated image with heat map on **cat2dog** dataset after 200 epoch

other samples especially when the features of cat such as ears or mouth are not obvious, the translated results are not satisfactory. Noticing that the heat map of the eighth sample has more weigh concentrating on the cat's face and there is no clear boundary between features such as ears and mouth. The same thing happened on the last sample, which we can't identify a clear face of dog on fake_B. The heat map of that sample is unclear and confused, whereas that of other samples has clear bound which depicts the feature of cat. We suppose that it is because of the features of the original image. Depending on the fur color and head position of cats, and image background, there are some features which are hard to be identified through the attention modular. When involving geometrical changes in translating images, a clear boundary between each feature is needed in order to identify each part and translate them into corresponding one.

In order to show the performance improvements of our modified model, we compared the results of our model with the those of CycleGAN model showed in figure5. Each

Figure 5. **cat2dog** results comparison of new model(upper) and CycleGAN(bottom)

model is trained over 200 epoch. The samples in first row are generated by our modified model and the second row are by original CycleGAN model. For the results of original model, it's not hard to notice that there is hardly no difference between real_A and fake_B, and hence it failed to translate from cat to dog even after 200 epoch. By analyzing the loss of original CycleGAN model, we found that it learned some patterns to reduce the overall loss, yet it didn't understand its task to translating the graphical features of the images.

Besides the experiment on **cat2dog** dataset, we also ran several experiments on other dataset because we want to ensure that our modified model can not only fulfill image to image translating tasks involving graphical changes but also have good performance on tasks involving texture or color transformation, such as converting horses to zebras. We have prior knowledge that the CycleGAN model is good at texture and color transformation. To justify our thoughts, we also use the CycleGAN model running on the same dataset for comparison.

4.4.2. Dataset: horse2zebra

Figure6 shows the generated image with heat map on **horse2zebra** dataset after 200 epoch. From this figure we could find out that the horses were indeed converted into zebras and the results are quite satisfactory. In most images the heat map emphasized the shapes of horses, focusing on changing that specific area into texture of zebras. However, there are still some limitations on this model. For example, in the seventh sample where there is a tree beside the horse, even though the weights are concentrated on the horse, the tree is also converted into texture of zebra. Such limitation also happens on CycleGAN according to our experiment results.

Figure7 is the comparison of generated images for **horse2zebra** dataset. The upper samples are generated using modified model and the bottom are from CycleGAN model. Generally, there are no much differences between results for both models, except for the color of background or texture, but such nuance won't affect overall performance.

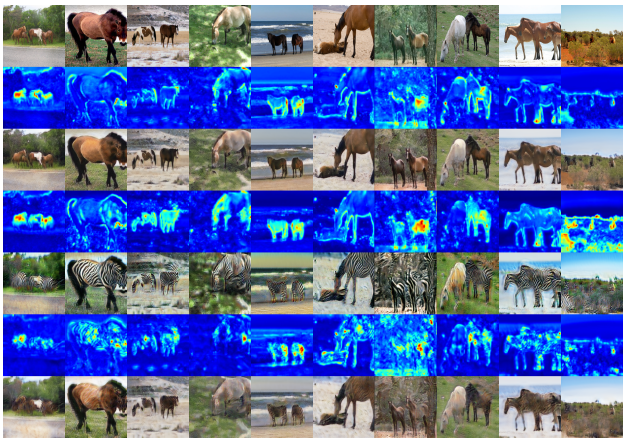


Figure 6. Generated image with heat map on **horse2zebra** dataset after 200 epoch

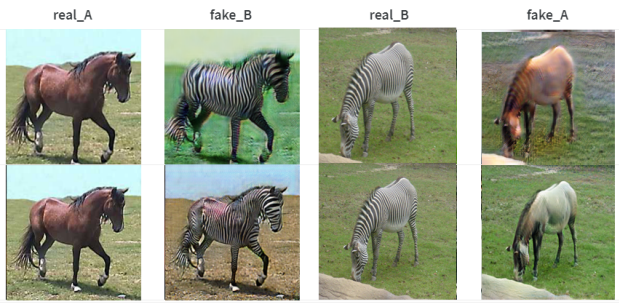


Figure 7. **horse2zebra** results comparison of new model(upper) and CycleGAN(bottom)

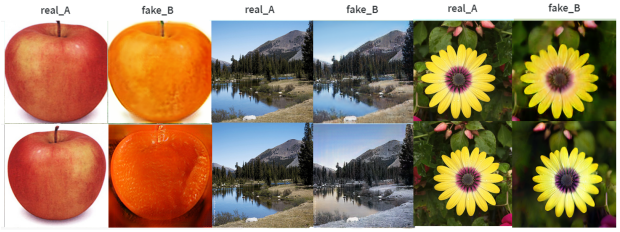


Figure 8. **apple2orange**, **summer2winter**, **iphone2dslr_flower** dataset results comparison of new model(upper) and CycleGAN(bottom)

Data set	Better than CycleGAN?
cat2dog	√
horse2zebra	√
apple2orange	√
summer2winter	×
iphone2dslr_flower	×

Table 3. Overall performance on different dataset

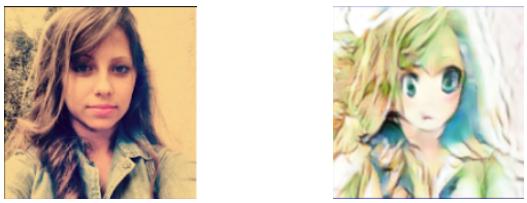
4.4.3. Results for other dataset

In order to check whether our modified model could outperform the original CycleGAN model on various dataset, we tested them with different dataset. The results for the rest experiments on dataset **apple2orange**, **summer2winter** and **iphone2dslr_flower** are showed in figure8. In these experiments we found that not all tasks involving color and texture transformation could be solved using the existing model.

The first two columns are for experiments on **apple2orange** dataset. In this task, the biggest problem is that neither our model nor CycleGAN model successfully remove the pedicel of the apple, which is absolutely not on oranges. We suggest that it is because this feature is relatively small that the attention module can't catch up this feature. The images in the middle two columns are for experiments on **summer2winter** dataset. In our results there is nothing changed between real_A and fake_B and hence our modfified model failed to do this task. The last two columns are for experiments on **iphone2dslr_flower** dataset. Different from previous experiments, this task aims to enhance the image qualities by improving the quality of photos taken by iPhone to the quality of photos taken by DSLR. Image enhancement is another kind of features for CycleGAN and we want to test that if our modified model could also do this kind of task. According to the results, the image translated by CycleGAN has higher quality and richer colors than ours. In summary, the overall results of these five datasets for both models are showed in table3.

4.4.4. Special experiment

We did an extra experiment on dataset **selfie2anime** due to the special feature of the network architectures which we adapted from UGATIT(J. Kim, 2019). According to their work, the UGATIT model could fulfill image style translation such as converting a realistic photo to anime. The author showed the results of anime-style transfer and the performance was really nice. Hence we ran the **selfie2anime** dataset on our modified model to check whether we could achieve the same performance as UGATIT.



(a) Real A (b) Fake B

Figure 9. **selfie2anime** dataset results

Figure9 shows the results of **selfie2anime** dataset. The overall style is anime though, there are some confused part in the generated image, such as the ambiguous boundary between hair and close, big and small eyes, and meaningless background. This outcome is below our expectation. We suggest that it might be due to the dataset we used in this experiment. We found that the distribution of selfie and anime are quite different. For example, in training examples of selfie, people wear glasses, use cellphone, make different poses, all of which have no corresponding items in anime. On the other hand, there are some characteristics in anime samples such as exaggerated head to body ratio which doesn't exist in realistic. Such discrepancy made it hard to do the selfie to anime translation, and we would only get good results on limited samples.

5. Related work

5.1. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) (I. Goodfellow, 2014a) have demonstrated remarkable performance in image generation (A. Radford, 2015a)(A. Radford, 2015b)(T. Karras, 2017), image editing (J. Zhu, 2016), image inpainting (S. Iizuka, 2017), image translation (Y. Choi, 2018)(X. Huang, 2018), and representation learning (M. F. Mathieu, 2016). The generator attempts to produce realistic pictures in order to deceive the discriminator, while the discriminator attempts to separate the generated image from the real image, which is made possible by GAN's concept of adversarial loss. The concept has recently gained widespread use in applications such as conditional picture creation (S. Reed, 2016) and other domains such as 3D data (J. Wu, 2016). The emergence of GANs permits translation from a source domain to a radically different destination domain provided unpaired training data (T. Wang, 2018).

5.2. Image-to-Image Translation

The concept of image-to-image translation can be traced back to Image Analogies (A. Hertzmann, 2001), in which Hertzmann used a single input-output training image pair to train a nonparametric texture model. Isola presented a unified framework for image-to-image translation based on conditional GANs, dubbed "pix2pix" (P. Isola, 2016). The "pix2pix" framework employs conditional generative adversarial networks to learn the mapping between the input and output images (I. Goodfellow, 2014b). CycleGAN's approach also makes use of this framework (**?**), since it is used to learn parametric translation from a dataset of input and output instances (J. Long, 2015). These concepts have been applied to a variety of techniques, including image generation from sketches (P. Sangkloy, 2017) or semantic layouts (L. Karacan, 2016). Recently, a high-resolution version of the "pix2pix" framework was developed (T. Wang, 2018). UNIT provides a shared workspace for unsupervised image translation (M. Liu, 2017). MUNIT decomposes images into domain-invariant content codes and domain-specific style codes (X. Huang, 2018). DRIT decomposes images into their constituent content and styles (H. Lee, 2018), enabling MUNIT to do many-to-many mapping. AGGAN enhances image translation performance by utilising an attention mechanism to discern foreground from background (Y. A. Mejjati, 2018). Although the idea of attention is novel and the experiment has a forward-looking directing impact, the indicator does not improve significantly.

5.3. Cycle Consistency

The concept of cycle consistency has gained more and more attention in data structure. Translators (Brislin, 1970) and translation machines (D. He, 2016) are rapidly embracing technological advances in their job of language translation, namely reverse translation and altering matching logic to test the correctness of translation, which may be improved by repeated verification. Visual tracking is constantly emphasising the consistency of attributes prior to and following the change (Z. Kalal, 2010). Cyclic consistency is being developed at a breakneck pace and has already produced remarkable achievements in the following areas: structural analysis under grasping motion state (C. Zach, 2010), high-precision 3D model matching (Q. Huang, 2013), adaptive dense semantic communities (T. Zhou, 2015), and deep-level estimation (C. Godard, 2017). It is worthwhile to learn a translation approach inspired by machine translation that can employ a similar purpose of capturing the invisible traits of unpaired images (Z. Yi, 2017).

5.4. Class Activation Map (CAM)

Class Activation Map (CAM) was proposed through utilising global average pooling in a CNN (B. Zhou, 2016). The CNN uses the CAM for a certain class to identify the image patches that may be used to make a distinction.

5.5. Normalization

Recent studies on neuronal style transfer have demonstrated that CNN feature statistics may be employed directly as descriptors of visual styles (L. A. Gatys, 2016). By normalising the image's extracted features, Instance Normalization (IN) eliminates stylistic variance. Recent researches, however, have discovered that when normalising images, Conditional Instance Normalization (CIN) (V. Dumoulin, 2016) and Batch-Instance Normalization (BIN) (H. Nam, 2018) are superior to IN alone.

6. Conclusions

In summary, the overall performance of our model is consistent with our expectations based on the results of those six experiments, though it still has many flaws and defects. In cat to dog transferring task, the model shows its ability to make geometrical translations in images. The attention module helps the model focus more on the features of source domain and target domain which made the geometrical change more precise. On texture and color transferring tasks, the model is also competent to carry out this kind of work, despite that the performance is not as good as CycleGAN on some certain dataset. It can be inferred that a model which performs good on tasks with geometrical changes doesn't represent that it can also perform good on tasks with texture and color changes. Additionally, we also tried to reproduce the results of selfie to anime tasks which is mentioned in the paper of UGATIT(J. Kim, 2019). However, we found it difficult to achieve the same performance as the authors did.

While our method produces acceptable outcomes in the great majority of circumstances, certain results are inconsistently satisfactory. In terms of contour judgement, our method cannot be completely accurate in identifying the attribution of contours with dominant chromatic aberration or invisible traits, since this is dependent on the threshold configuration of the discriminator for contour judgement. This is more dominantly reflected in the process of converting a horse into a zebra, which can be accomplished entirely even while the horse is partially concealed. The issue is that in such a case, if a person performs a similar movement, our method is likely to misidentify it and converted into a zebra man. To solve this issue, further effort may need optimising the discriminator's attention in order to properly locate object outlines by the incorporation of additional external features.

Another noticeable flaw is that it is difficult for the generator to convert the original data set with obvious morphological changes to the target data set in order to accomplish an entirely perfect conversion, which can also serve as a direction for future research. The generator is not capable of concealing or erasing morphological traits that should not be present in the generated target yet are. A more typical case occurs when apples are converted to oranges, as apples include pedicels in their natural state, which oranges normally lack, yet our method makes it impossible to obscure

or eliminate this trait.

References

EECS Berkeley, 2021. URL http://efrosgans.eecs.berkeley.edu/cyclegan/datasets/.

A. Almahairi, S. Rajeshwar, A. Sordoni P. Bachman A. Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *Proceedings of the 35th International Conference on Machine Learning*, 2018. URL https://proceedings.mlr.press/v80/almahairi18a.html.

A. Hertzmann, C. E. Jacobs, N. Oliver B. Curless D. H. Salesin. Image analogies. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001. URL https://doi.org/10.1145/383259.383295.

A. Makhzani, J. Shlens, N. Jaitly I. Goodfellow B. Frey. Adversarial autoencoders. *ICLR*, 2016. URL https://arxiv.org/abs/1511.05644.

A. Radford, L. Metz, S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015a. URL https://arxiv.org/abs/1511.06434.

A. Radford, L. Metz, S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015b. URL https://arxiv.org/abs/1511.06434.

Asirra, Petfinder. *Microsoft Research*, 2017. URL https://www.microsoft.com/en-us/download/details.aspx?id=54765.

B. Zhou, A. Khosla, A. Lapedriza A. Oliva A. Torralba. Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html.

Brislin, R. W. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1970. URL https://doi.org/10.1177/135910457000100301.

C. Godard, O. M. Aodha, G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1609.03677.

C. Zach, M. Klopschitz, M. Pollefeys. Disambiguating visual relations using loop constraints. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5539801.

D. He, Y. Xia, T. Qin L. Wang N. Yu T. Liu W. Ma. Dual learning for machine translation. *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/file/5b69b9cb83065d403869739ae7f0995e-Paper.pdf.

G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 2006. URL https://www.science.org/doi/10.1126/science.1127647.

Goodfellow, I. Generative adversarial networks. *NIPS 2016 Tutorial*, 2016. URL https://arxiv.org/abs/1701.00160.

H. Lee, H. Tseng, J. Huang M. Singh M. Yang. Diverse image-to-image translation via disentangled representations. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Hsin-Ying_Lee_Diverse_Image-to-Image_Translation_ECCV_2018_paper.html.

H. Nam, H. Kim. Batch-instance normalization for adaptively style-invariant neural networks. *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/file/018b59ce1fd616d874afad0f44ba338d-Paper.pdf.

I. Goodfellow, J. Pouget-Abadie. Generative adversarial nets. *arXiv preprint arXiv:1406.2661*, 2014a. URL https://arxiv.org/abs/1406.2661.

I. Goodfellow, J. Pouget-Abadie, M. Mirza B. Xu D. Warde-Farley S. Ozair A. Courville Y. Bengio. Generative adversarial nets. *NIPS*, 2014b. URL https://arxiv.org/abs/1406.2661.

J. Kim, M. Kim, H. Kang K. Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv:1907.10830*, 2019. URL https://arxiv.org/abs/1907.10830.

J. L. Ba, J. R. Kiros, G. E. Hinton. Layer normalizations. *arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html.

J. Wu, C. Zhang, T. Xue B. Freeman J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/file/44f683a84163b3523afe57c2e008bc8c-Paper.pdf.

J. Zhu, P. Krähenbühl, E. Shechtman A. A. Efros. Generative visual manipulation on the natural image manifold. *Computer Vision – ECCV 2016*, 2016. URL https://doi.org/10.1007/978-3-319-46454-1_36.

J. Zhu, T. Park, P. Isola. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1703.10593v6.

K. Xu, J. Ba, R. Kiros K. Cho A. Courville R. Salakhudinov R. Zemel Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2015. URL https://proceedings.mlr.press/v37/xuc15.html.

L. A. Gatys, A. S. Ecker, M. Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html.

L. Karacan, Z. Akata, A. Erdem E. Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv:1612.00215*, 2016. URL https://arxiv.org/abs/1612.00215.

M. F. Mathieu, J. Zhao, A. Ramesh P. Sprechmann Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf.

M. Liu, T. Breuel, J. Kautz. Unsupervised image-to-image translation networks. *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/file/dc6a6489640ca02b0d42dabeb8e46bb7-Paper.pdf.

P. Isola, J. Zhu, T. Zhou. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. URL https://arxiv.org/abs/1611.07004.

P. Sangkloy, J. Lu, C. Fang F. Yu J. Hays. Scribbler: Controlling deep image synthesis with sketch and color. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1612.00835v2.

Q. Huang, L. Guibas. Consistent shape maps via semidefinite programming. *Computer graphics forum*, 2013. URL https://doi.org/10.1111/cgf.12184.

S. Iizuka, E. Simo-Serra, H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 2017. URL https://doi.org/10.1145/3072959.3073659.

S. Reed, Z. Akata, X. Yan L. Logeswaran B. Schiele H. Lee. Generative adversarial text to image synthesis. *Proceedings of The 33rd International Conference on Machine Learning*, 2016. URL https://proceedings.mlr.press/v48/reed16.html.

Saxena, P. Cycle generative adversarial network (cyclegan), 2021. URL https://www.geeksforgeeks.org/cycle-generative-adversarial-network-cyclegan-2/.

T. Karras, T. Aila, S. Laine J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017. URL https://arxiv.org/abs/1710.10196.

T. Wang, M. Liu, J. Zhu A. Tao J. Kautz B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1711.11585v2.

T. Zhou, Y. J. Lee, S. X. Yu A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Zhou_FlowWeb_Joint_Image_2015_CVPR_paper.html.

V. Dumoulin, J. Shlens, M. Kudlur. A learned representation for artistic style. *arXiv:1610.07629*, 2016. URL https://arxiv.org/abs/1610.07629.

X. Huang, S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1703.06868v2.

X. Huang, M. Liu, S. Belongie J. Kautz. Multimodal unsupervised image-to-image translation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1804.04732v2.

X. Mao, Q. Li, H. Xie R. Y.K. Lau Z. Wang S. P. Smolley. Least squares generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1611.04076.

Y. A. Mejjati, C. Richardt, J. Tompkin D. Cosker K. I. Kim. Unsupervised attention-guided image-to-image translation. *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/file/4e87337f366f72daa424dae11df0538c-Paper.pdf.

Y. Choi, M. Choi, M. Kim J. Ha S. Kim J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://arxiv.org/abs/1711.09020v3.

Z. Kalal, K. Mikolajczyk, J. Matas. Forward-backward error: Automatic detection of tracking failures. *2010 20th International Conference on Pattern Recognition*, 2010. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5596017.

Z. Yi, H. Zhang, P. Tan M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1704.02510v4.