Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/18/20230364

# Unpaired image neural style transfer based on Non-Local-Attention-Cycle-Consistent adversarial network

**Hao Sun**[1,*]

[1]The department of Computer Science, Simon Fraser University, Burnaby, V5A 1S6, Canada

[*]hsa117@sfu.ca

**Abstract.** Existing image translation methods already enable style transfer on unpaired data. Although these methods have yielded satisfactory results, they still result in changing the background while changing the object. One reason is that when using convolutional neural networks, global information is lost as the number of network layers increases, and the absence of an effective sensory field leads to the failure to generate high-quality results. This paper proposed a Non-Local-Attention-Cycle-Consistent Adversarial Networks for unpaired images style transfer. The no-local-attention can quickly capture long-range dependencies, better extracts global information, ensures effective focus on the foreground while preserving the background, and can be easily embedded into the current network architecture. Experiments are conducted on neural style transfer task with public dataset, this model can obtain the better result than CycleGAN. It allows better attention to structural features rather than just textural features. It can reconstruct some of the content lost by CycleGAN. Recent research has also demonstrated that the optimizer has an impact on the performance of the network. This paper applies the Nadam optimizer and find that this improves training process.

**Keywords:** No-local Attention, CycleGAN, Style Transfer.

## 1. Introduction

Many painters have spent much time trying to imitate the style of famous artists. These artists have acquired the ability to create one-of-a-kind visual experiences by constructing a complex interaction between an image's content and style. In recent years, Artificial Intelligence has also made significant progress in remapping the content of a given image with a stylized image. This issue is more widely defined as a neural style transfer, which relates to a way of transforming the style of an input image into a method that specifies.

The first method to achieve style transfer is the Image Analogies [1], which is simple multiscale autoregression. The limitation of this model is that it can only extract the underlying features and miss the high order features, also it has the low speed to process the different types of images, such methods have developed slowly over the years. Convolutional Neural Networks (CNNs)-based supervised learning algorithms have gained popularity in recent years with the growth of neural network algorithms, however the low quality of the generated pictures is a result of the small size of the manually constructed libraries of paired training images. Until the advent of Generative Adversarial Networks (GANs), there was significant progress in this direction [2]. The field of neural style transfer is progressively using several techniques based on GAN concepts. A such method is pix2pix [3], which proposes a common framework to solve the challenges of neural style transfer by combining adversarial loss with L1 loss and using supervised learning for training based on cGANs [4]. Unfortunately getting matched training data may be costly and complex. It may be more difficult for graphical tasks such as style transfer because the required output is very complex and usually requires artistic creation. Therefore, unsupervised learning-based image style transformation frameworks are proposed. Rosales provides a probabilistic inference and learning method that is computationally effective for learning the rendering style and predicting the most likely output image [5]. The important information of the input and generate images is preserved via cyclic consistent loss by Cycle-consistent Generative Adversarial Network (CycleGAN), DiscoGAN and DualGAN [6-8]. This enables neural style transfer in the unpaired data.

CycleGAN, a typical neural style transfer method in the GAN family, includes two basic GAN structures to generate and train by unpaired data. But the author also pointed out the shortcomings, due to the relationship of the generator model, it is more inclined to change the appearance of the input image rather than the shape. And when the input image contains objects that are not contained in the input domain during training, the generator's mapping will produce various changes. In short, it will change the background while changing the object. The generator uses a residual network, but it still has a limited field of perception. Even though the theoretical field of perception becomes larger as the number of layers and convolution kernels increases, the effective field of perception (the field of perception that works) of the features is much smaller than the theoretical field of perception according to Luo's paper [9]. Moreover, when the number of network layers increases, the global information of features will be lost, which will have a negative impact on style transfer. According to Wang's article, convolutional networks have three problems in extracting global information, the first one is that capturing long range features relies on networks that need to accumulate many layers, resulting in too low learning efficiency, the second one is that careful design of modules and gradients is needed because the networks need to accumulate very deep, and the last one is that convolutional or temporal local operations are difficult when messages need to be passed back and forth between relatively distant locations [10].

This study proposes Non-Local-Attention-Cycle-Consistent Adversarial Networks (NLACycleGAN) to deal with the problems, which introduce a non-local attention into CycleGAN. The non-local attention module in deep neural networks is a useful and adaptable element for long-range relationships [10]. There are some advantages of using non-local attention, it can directly calculate the relationship between two spatial locations to quickly capture long range dependencies, and it has highly computationally efficient, requiring fewer stacked layers to achieve the same result, the important is the input scale and output scale can be guaranteed to be constant, and can be easily embedded in the current network architecture. The most important of this model is to better extract the global information to ensure effective focus on the foreground while preserving the background.

In addition to Non-Local-Attention, this paper also changes the optimizer. The article showed that the Nadam is a faster, more powerful learning algorithm [11]. It places a tighter restriction on the learning rate and directly affects the gradient update.

## 2. Method

### *2.1. Dataset description and preprocessing*

In this paper, the datasets used monet2photo [12]. This dataset includes two classes, namely art images and real photos. Using the categories landscape and landscapephotography, the actual pictures were downloaded from Flickr, while the art pictures were downloaded from Wikiart [13-14]. The size of each class was 1074 (Monet) and 6853 (Photographs).

In terms of data preprocessing, it can be divided into two parts. First, the images were resized to 286 x 286 pixels since the dataset has images with different size. Resizing make sure they are the same size so that it guaranteed network stability when these image pass through the network as input. Second, to prevent overfitting to a set of specific features, these images randomly cropped, and the crop size is 256 x 256 pixels. Figure 1 shows the visualization of some images after processing.

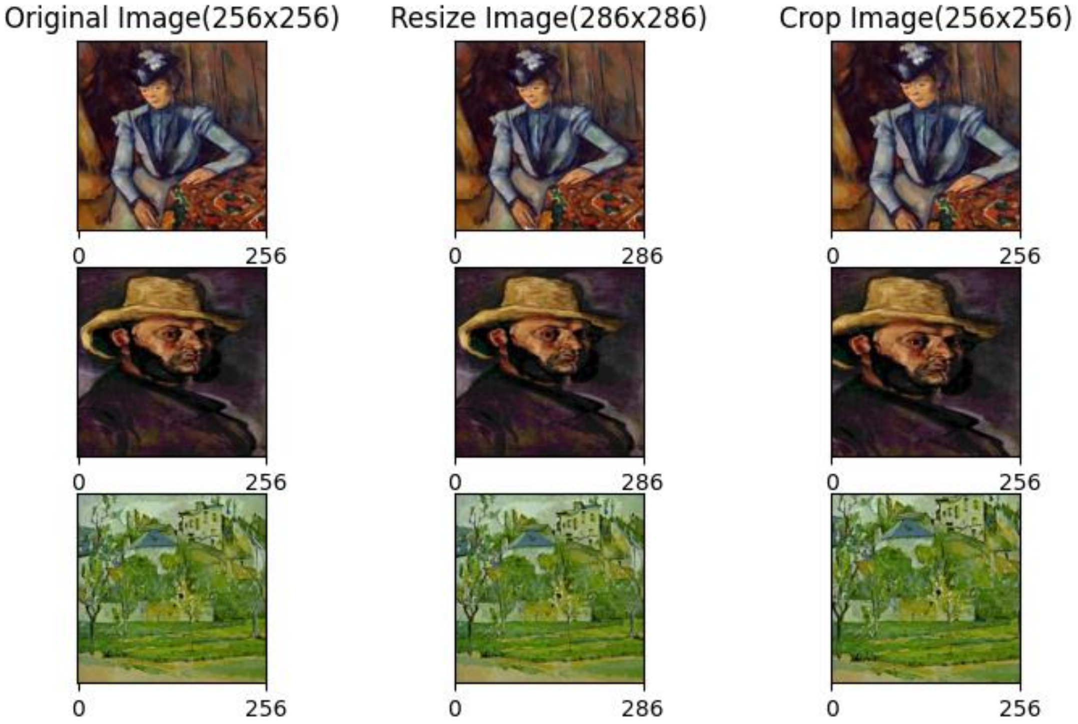

**Figure 1.** Visualizations of Images after processing.

### *2.2. Proposed approach*

In this paper, the based model is CycleGAN, and the non-local attention is introduced in this network. For basic GAN, it consists of two competing modules: a generator G and a discriminator D. The generator G and the discriminator D, which are iteratively trained iteratively, compete. For CycleGAN, it consists of two adversarial discriminators $D_X$ and $D_Y$ as well as two generator G: $G_{X->Y}$ and F:$G_{Y->X}$. Generator G attempts to trick the discriminator $D_Y$ by causing the image of the image domain X to transit through this network in order to create the image of the image domain Y. Additionally, $D_Y$ works to improve itself so that it can determine if the sample is a picture that was produced or one that was created using actual data. The basic architecture of the CycleGAN can be found in Figure 2.

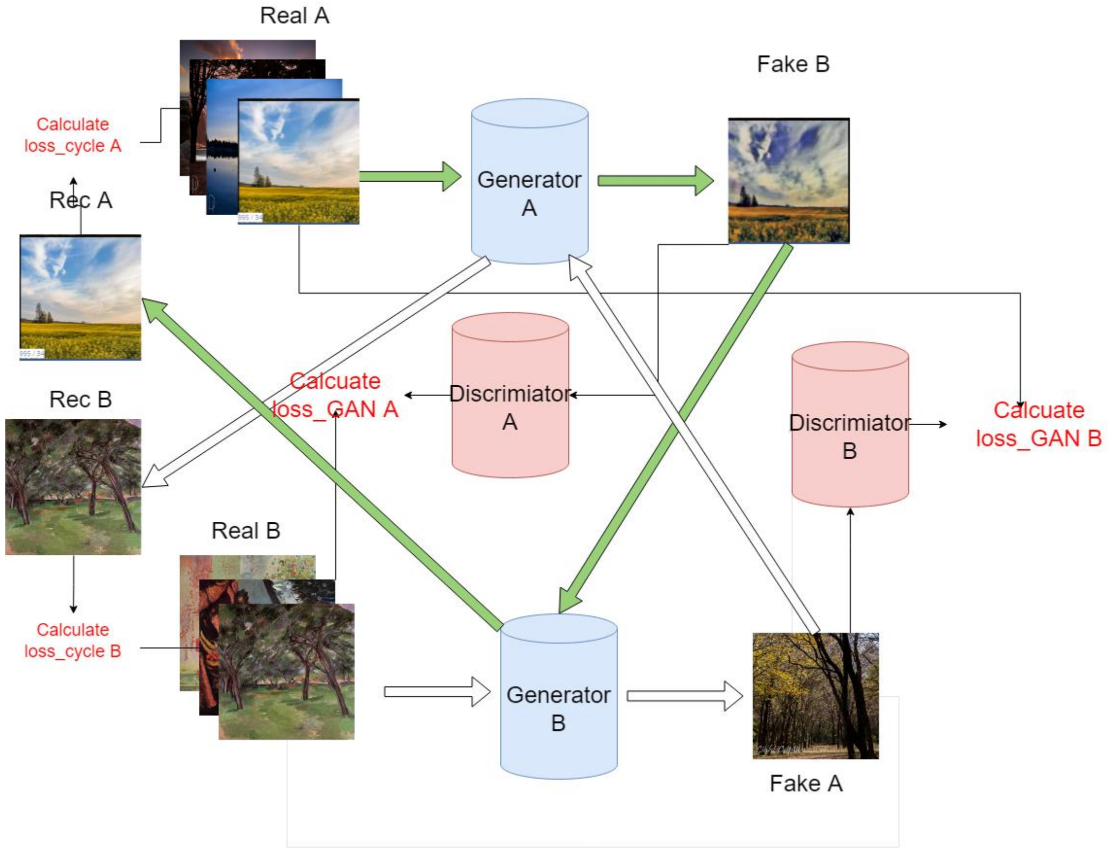

**Figure 2.** The basic CycleGAN model.

Since the generator is build using Resnet model in this study, it based on convolutional layers. Convolution only focuses on information within the local neighborhood, so using convolution layers alone is computationally inefficient. And as the depth of the convolution layer increases, the global information will be lost. It causes the details lost when the image translation. To solve this issue, this paper introduces the non-local attention, which is adapted by self-attention, into the CycleGAN framework.

Non-local is an easy to integrate. And its base idea is refining information for a feature map. Its universal format is $y_i = 1/C(x) \sum f(x_i, x_j) g(x_j)$, where x is the feature map as input, i is the output index, $g(x_j) = W_g x_j$, it is a linear embedding, $W_g$ is the weight matrix to be learned, which can be achieved by 1x1 convolution on the space. The important function is $f(x_i, x_j)$ , it calculates the similarity of i and j. In this paper, f function uses Embedded Gaussian method. It is $f(x_i, x_j) = e^{(\theta(x_i)^T \phi(x_j))}$, where $x_i$ represents the information of the current position of interest, and $x_j$ represents the global information. After implementing non-local network, this paper needs to transform it to non-local block to replace the residual network block. So, this paper defined a non-local block $z_i = W_z y_i + x_i$, where $x_i$ is same like a residual connection. As a result, it could produce feature maps with the same size as the input feature maps. It easy to be inserted into the CycleGAN framework. And it has been applied to the generator, which are trained by least-squares loss of the adversarial loss. For other loss function, it same as the CycleGAN. Cycle consistency loss makes sure that the content of the output picture of the generator is the same and only differs in style from the input image. Identity loss make sure the tone of the generator's output picture matches that of the input image. Figure 3 presents the architecture of the generator in No-local CycleGAN.

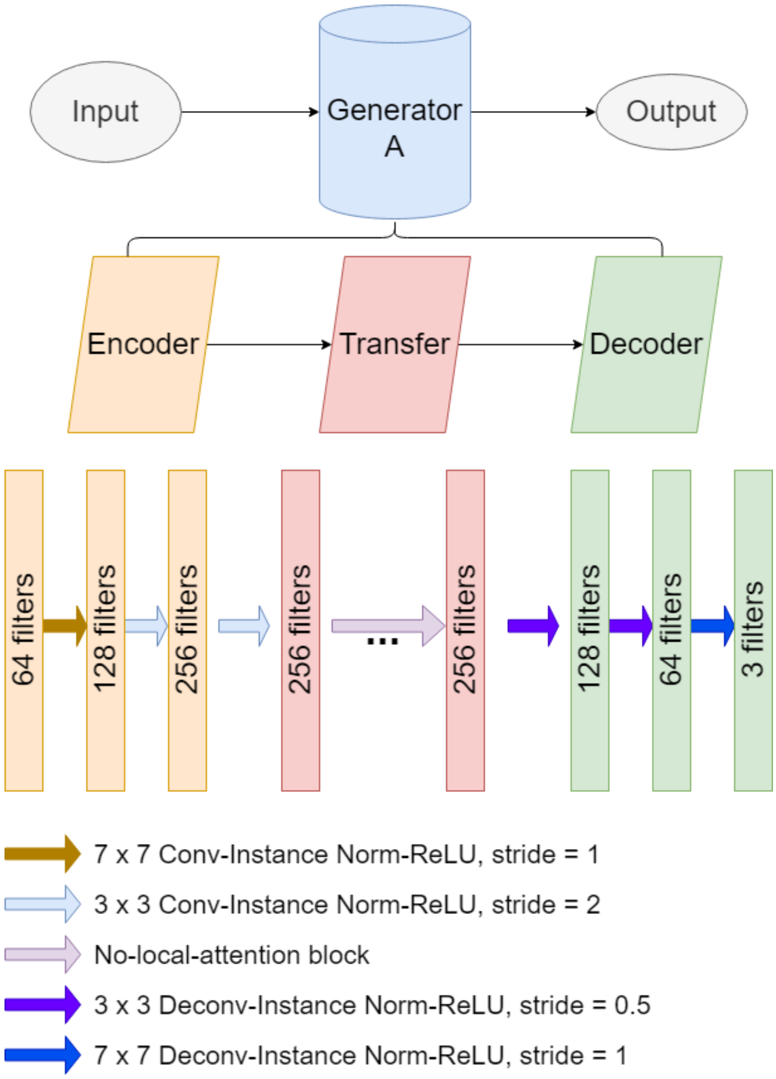
Input
Generator A
Output
Encoder
Transfer
Decoder
64 filters
128 filters
256 filters
256 filters
...
256 filters
128 filters
64 filters
3 filters
7 x 7 Conv-Instance Norm-ReLU, stride = 1
3 x 3 Conv-Instance Norm-ReLU, stride = 2
No-local-attention block
3 x 3 Deconv-Instance Norm-ReLU, stride = 0.5
7 x 7 Deconv-Instance Norm-ReLU, stride = 1

**Figure 3.** The architecture of the generator in No-local-attention CycleGAN.

### *2.3. Implementation Details*

All the models designed in the Pytorch. Instance normalization is applied by default to the layers of the generator and discriminator. And the model use Nadam optimizer. For both the discriminator and the generator, the learning rate is 0.0002. Since the learning rate decays linearly in later epochs beyond the first 100, the epoch is 150.

## 3. Results and discussion

### *3.1. Analysis of Optimizer.*

In the original CycleGAN model, it uses Adam optimizer, which is effective control of learning rate step and gradient direction by first-order momentum and second-order momentum. And it will not have the global optimal solution, because it does not converge. So, in this paper, it changes to Nadam, it imposes a more stringent limit on the learning rate and directly influences the gradient update. The compare result is shown in Figure 4 and Figure 5. Because the Nadam is like Adam with Nesterov momentum term. These two optimizers have the same formula for the part of the calculation of momentum, both are $m_t = \mu * m_{t-1} + (1 - \mu) * g_t$ and $n_t = \nu * n_{t-1} + (1 - \nu) * g_t^2$. But they are different in the correction for momentum, for Adam, its $\widehat{\mathrm{m}_\mathrm{t}} = \frac{\mathrm{m}_\mathrm{t}}{1-\mu^\mathrm{t}}$, $\widehat{\mathrm{n}_\mathrm{t}} = \frac{\mathrm{n}_\mathrm{t}}{1-\nu^\mathrm{t}}$,for Nadam, its $\widehat{\mathrm{m}_\mathrm{t}} = \frac{\mathrm{m}_\mathrm{t}}{1-\prod_{\mathrm{i}=1}^{\mathrm{t}+1}\mu_i}$, $\widehat{n_t} = \frac{n_t}{1-\nu^t}\overline{m_t} = (1-\mu_t) * \widehat{g_t} + \mu_t + 1 * \widehat{m_t}$, it does not calculate the direction of the gradient at the current position, but the direction of descent at that time if a step is taken according to the accumulated momentum is $\widehat{g_t}$.

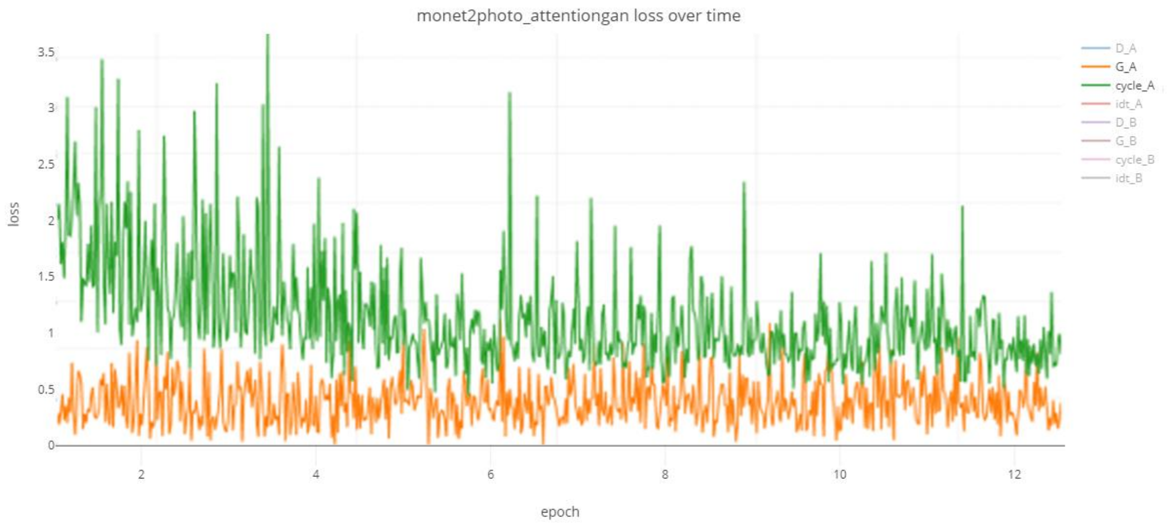

**Figure 4.** The loss value of using Adam.

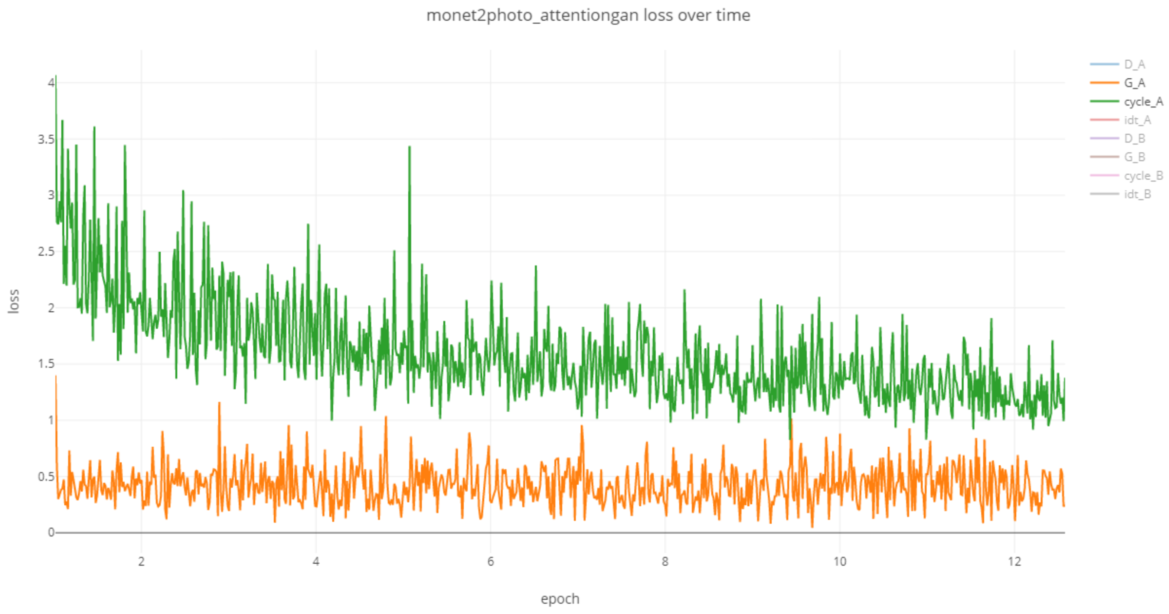

**Figure 5.** The loss value of using Nadam.

### *3.2. Result of Style Transfer*

Figure 6 shows some results of style transfer images for test data in dataset. It observes that the proposed No-local-Attention CycleGAN produces clearer and more diverse results than CycleGAN when generating images with complex geometric or structural patterns. For other class, the No-local-Attention CycleGAN shows less advantage with CycleGAN. The generator of the model outputs an attention maps instead of restoring the full image. In the last row of Figure 6, it can be observed that different inputs produce different attention maps. By handling the geometric differences between the source and target domains, this technique enables the generator to concentrate solely on the areas of the picture that have been recognized as producing the new expressions. This reinforces the motivation's initial validity. The technique also retains extra or undesirable components of the picture. The reason is that self-attention receives more evidence and allows freer selection of conditions with larger feature mappings, and it demonstrates that the attention mechanism provides the generator and discriminator with greater ability to directly model long-term dependencies in feature mappings. However, it has no significant improvement over the CycleGAN when dealing with images that are distinguished by texture rather than structure. The reason is that the no-local-attention is complementary to the convolution for global level. If the input only has the simple texture, its work as the local convolution.

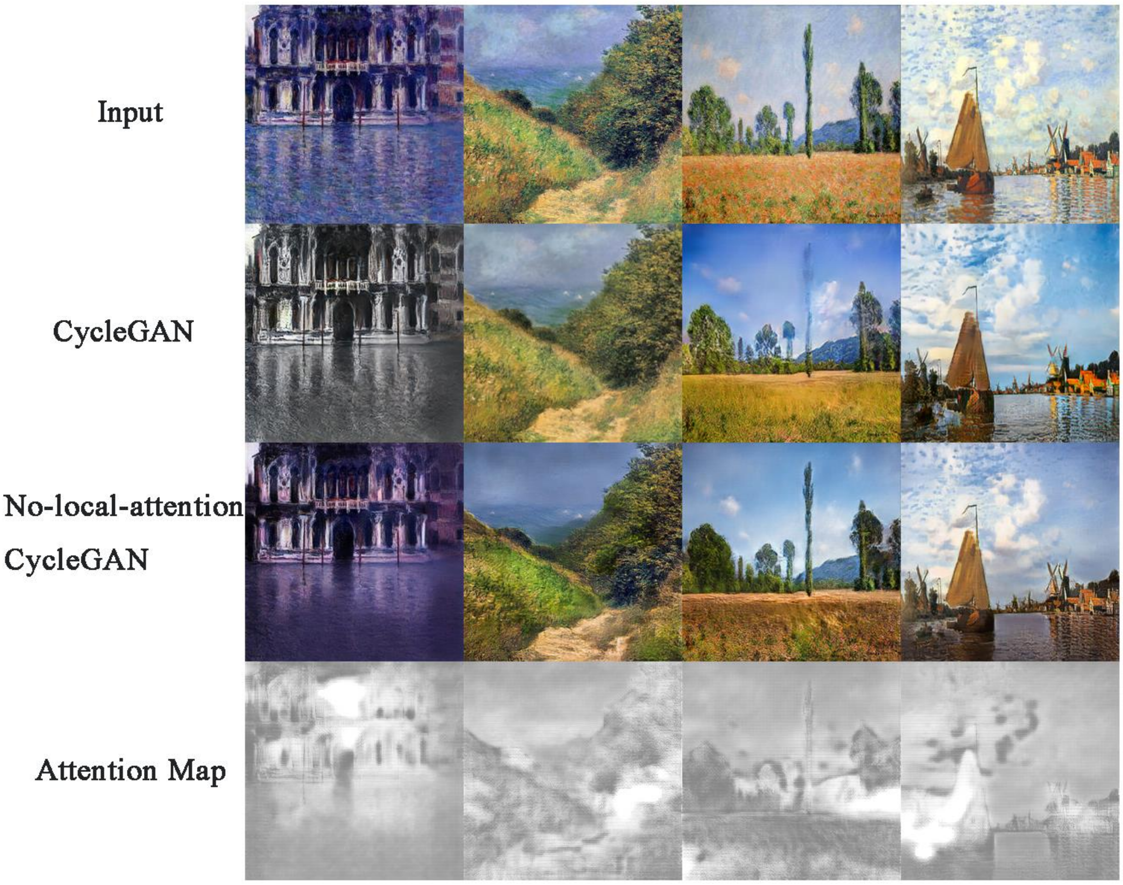

**Figure 6.** The result of CycleGAN and No-local-attention CycleGAN.

## 4. Conclusion

This paper proposed No-Local-Attention Cycle-consistent Generative Adversarial Networks (NLACycleGANs) for unpaired neural style transfer, which introduce a no-local-attention mechanism into the CycleGAN framework. It can produce attention maps, gather global features, and identify the source and target domains' most discriminating content. Then the attention maps are combined with the input image to generate a high-quality target image. In addition, this paper also shows that the Nadam optimizer can speed up the convergence rate and learning rate. Numerous experimental findings on the neural style transfer task show that the proposed model is capable of producing results that are more trustworthy than CycleGAN. The model has shortcomings in dealing with more textured inputs, and the network design can be improved to solve this problem in feature studies.

## References

[1] Hertzmann A et al. 2001 Image analogies Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’01 New York New York

[2] Goodfellow IJ et al. 2014 Generative Adversarial Networks arXiv [statML] doi:1048550/ARXIV14062661 http://arxivorg/abs/14062661

[3] Isola P et al. 2016 Image-to-image translation with conditional adversarial networks arXiv [csCV] doi:1048550/ARXIV161107004 http://arxivorg/abs/161107004

[4] Mirza M et al.2014 Conditional generative Adversarial Nets arXiv [csLG] doi:1048550/ARXIV14111784 http://arxivorg/abs/14111784

[5] Resales A F 2003 Unsupervised image translation In: Proceedings Ninth IEEE International Conference on Computer Vision IEEE p 472–478 vol 1

[6] Zhu J-Y et al. 2017 Unpaired image-to-image translation using cycle-consistent adversarial networks arXiv [csCV] doi:1048550/ARXIV170310593 http://arxivorg/abs/170310593

[7] Kim T et al. 2017 Learning to discover cross-domain relations with generative adversarial networks arXiv [csCV] doi:1048550/ARXIV170305192 http://arxivorg/abs/170305192

[8] Yi Z et al. 2017 DualGAN: Unsupervised dual learning for image-to-image translation arXiv [csCV] doi:1048550/ARXIV170402510 [accessed 2022 Sep 29] http://arxivorg/abs/170402510

[9] Luo W et al. 2017 Understanding the effective receptive field in deep convolutional neural networks arXiv [csCV] doi:1048550/ARXIV170104128 http://arxivorg/abs/170104128

[10] Wang X et al. 2017 Non-local Neural Networks arXiv [csCV] doi:1048550/ARXIV171107971 http://arxivorg/abs/171107971

[11] Dozat T 2015 Incorporating Nesterov momentum into Adam Stanfordedu https://cs229stanfordedu/proj2015/054_reportpdf

[12] Index of /~taesung_park/CycleGAN/datasets Berkeleyedu https://peopleeecsberkeleyedu/~taesung_park/CycleGAN/datasets/

[13] Christiansen L C 2014 Find your inspiration: Finding your balance of health and fitness Lawton OK: Penguin International Publishing

[14] Wikiartorg 2022 visual art encyclopedia https://www.wikiart.org/