Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/51/20241192

Ancient Chinese painting style transfer based on CycleGAN

Jianwei Bai

Department of Mathematics, Hong Kong Baptist University, Hong Kong, 999077, China

23468181@life.hkbu.edu.hk

Abstract. Style transfer aims to alter the visual aesthetic of images by giving them a different artistic style. With the rapid advancement of deep learning, style transfer tasks have made significant progress, introducing new perspectives and innovative potentials within the realm of image processing. This study seeks to explore style transfer methods based on Cycle-Consistent Adversarial Networks (CycleGAN), enabling contemporary landscape photographs to take on the form of ancient Chinese paintings. This endeavor opens up fresh possibilities for artistic creation, image editing and design applications. The research encompasses an exposition of the process involved in constructing the CycleGAN model, alongside presenting research findings. Furthermore, it delves into the discussion of crucial techniques employed during the model training process, specifically the utilization of cycle consistency loss in configuring the loss functions. Lastly, this study ventures into future research directions, including strategies for further enhancing the performance and expanding the application scope of this style transfer model.

Keywords: Style Transfer, Ancient Chinese Painting, CycleGAN, Deep Learning.

1. Introduction

For millennia, individuals have been captivated by the realm of visual arts, wherein artists have crafted numerous awe-inspiring creations that have endured through time [1]. Historically, the process of reimagining images in specific artistic styles demanded profound artistic skills and consumed extensive periods. However, since the mid-1990s, the theoretical underpinnings of these remarkable artistic endeavors have not only engrossed the attention of artists but also garnered the interest of numerous computer science researchers. This has catalyzed an abundance of research and technological exploration focused on the automated transformation of real images into works of art [2].

In the early stages of style transfer, the prevalent approach involved the creation of dedicated models tailored to specific stylistic images. However, this method had a substantial limitation in that a single program could only perform style transfer for a particular style or scene, resulting in significant constraints in practical applications [3].

In recent years, the advancements in deep learning have opened up numerous novel possibilities in the realm of style transfer. Gatys and his colleagues achieved successful style transfer by employing convolutional neural networks to extract features related to image content and style [4]. Subsequently, methods based on Generative Adversarial Networks (GANs) not only achieved remarkable success in tasks like image generation and transformation but also began to find applications in style transfer tasks

DOI: 10.54254/2755-2721/51/20241192

[5]. A pivotal milestone came with the emergence of Cycle-Consistent Adversarial Networks (CycleGAN), which not only expanded the scope of GANs applications but also triggered a revolutionary transformation within the field of style transfer [6].

The primary objective of this study is to transform contemporary landscape photographs into the style of ancient Chinese paintings. However, traditional style transfer methods face substantial challenges in achieving this goal due to the stark disparity in content between ancient Chinese paintings, which often depict Chinese landscapes and figures, and modern landscape photographs, which predominantly feature global natural scenery and urban streetscapes. Traditional methods necessitate strict content matching, adding complexity to the data preparation process [7].

In contrast, the CycleGAN model presents a notable advantage in that it does not require content matching during the data preparation stage. This means it can utilize modern photos and ancient paintings with varying scenes and themes without the need for manually creating paired datasets. This significantly reduces the complexity of data preparation and diminishes the reliance on content matching, thereby broadening the spectrum of style transfer possibilities.

CycleGAN, being founded on GANs, possesses the capability to generate high-quality stylized images, resulting in transformation outcomes that are more lifelike and visually appealing. In this study, the achievement of style transfer between modern photographs and ancient Chinese paintings will be realized through the training of the CycleGAN model.

2. Method

2.1. Dataset

The dataset used in this project was acquired from the Kaggle website, consisting of modern images (train: test = 6287: 751) and ancient images (train: test = 562: 263), all in jpg format [8]. The modern-style images predominantly comprise landscape photographs captured with cameras, encompassing various natural sceneries and urban landscapes. In contrast, the ancient-style images basically encompass a wide range of common ancient Chinese paintings, including diverse forms such as ink painting, gongbi painting, freehand painting, border painting, and printmaking. The primary subjects within the paintings encompass figures, animals, landscapes, etc. Representative images are demonstrated in Figure 1.

The diversity of image styles within the dataset is crucial for training the model to adapt to various themes and scenes, which enhances the model's generalization capabilities, enabling it to excel in the task of style transformation for modern photos.

Figure 1. Examples of modern pictures (upper) and ancient Chinese paintings (lower) [8].

DOI: 10.54254/2755-2721/51/20241192

In the data preprocessing phase, the initial step comprised resizing all images to dimensions of 256x256 pixels. Subsequently, data augmentation was applied with a 50% probability of random flipping to enhance dataset diversity. Following this, the images underwent normalization to ensure uniformity in input data, thereby enhancing the stability of the model training process. Lastly, the image data was converted into tensor format.

2.2. Architecture of CycleGAN

CycleGAN is an unsupervised deep learning model, with its fundamental concept revolving around the utilization of two sets of neural networks: two Generators and two Discriminators, to facilitate style transformation between images [6,9]. The primary role of the Generator is to transfer the style of one image into another, while the Discriminator's objective is to discern between generated images and authentic ones.

The functionality implemented involves taking an input image into the Generator, where the Generator strives to create a new image that visually aligns with the target style. Through training, the goal is to generate images of sufficient realism, making it as challenging as possible for the Discriminator to distinguish them from real images. The Discriminator, through its training, becomes more adept at distinguishing between generated images and original ones, thereby encouraging the Generator to produce higher-quality images. Throughout the training process, these two components collaborate to achieve images' style transfer goal [10].

2.2.1. Architecture of Discriminator

In the construction of the Discriminator, a five-layer convolutional neural network architecture was employed to determine whether an input image was generated by the Generator or originated from real images. The input image dimension was set at 3x256x256 pixels. The first four convolutional layers were dedicated to feature extraction, with each layer progressively increasing the number of channels, following a pattern of features = [64, 128, 256, 512]. This augmentation in network depth aimed to enhance the receptive field and feature extraction capacity of the model. With each added layer, the network could learn higher-level abstract features, facilitating its ability to better discern input images.

The final convolutional layer served as the output layer, producing a single-channel result. During the forward propagation process of the model, the resulting feature values were mapped to the range [0,1] using a Sigmoid function, rendering the final output of the discriminator in the form of a probability distribution, allowing for binary classification.

Within the construction of the convolutional layers, Instance Normalization was incorporated. Its purpose lies in ensuring that the mean and variance within each channel closely approximate the desired normalized values. This normalization process fosters greater consistency in feature distributions across different channels, thereby enhancing network training stability and generalization capability.

Simultaneously, the LeakyReLU activation function was employed to mitigate gradient vanishing issues. Given that the Discriminator plays a role in providing information to aid the Generator's improvement, LeakyReLU activation was set to ensure that gradients do not vanish even when the input values are negative. This, in turn, bolsters the stability of model training.

2.2.2. Architecture of Generator

The Generator's primary structure comprises convolutional layers for downsampling, residual layers, and convolutional layers for upsampling, facilitating the transformation of an input image with dimensions of 3x256x256 pixels into an image with an altered style.

Beginning with the upsampling layers, feature extraction is initiated by progressively increasing the number of channels to capture higher-level features. The construction of the convolutional layers in this phase also integrates Instance Normalization and ReLU activation functions. Subsequently, the network's depth is augmented through the incorporation of residual blocks, aiding the model in learning more intricate feature representations and generating more realistic images.

DOI: 10.54254/2755-2721/51/20241192

The inclusion of residual blocks is essential to mitigate potential gradient vanishing issues that can arise as the network's depth increases, thereby facilitating smoother convergence. Each residual layer is comprised of two convolutional layers, which serve to perform nonlinear feature transformations and identity mapping. This design preserves and enhances fine-grained image details, alleviating information loss associated with deep networks.

Following this, the upsampling phase employs symmetrically designed transposed convolutions, mirroring the structure of the downsampling section. This gradually increases the spatial dimensions of the feature maps, enabling the generation of higher-resolution images. The final convolutional layer configuration is tasked with mapping the features back to the original image channel count, thus restoring the spatial dimensions of the generated image to ensure it matches the resolution of the original image.

2.2.3. Loss Function and Hyperparameter Setting

During the backward propagation process in model training, distinct loss functions were assigned to the Discriminator and Generator.

For the Discriminator's loss function configuration, the process begins by generating images using the Generator. Subsequently, both the original and generated images are individually assessed by the Discriminator to produce classification outcomes. The desired Discriminator performance is centered on accurate classification, aiming for a classification result as close to 1 as possible for input original images and as close to 0 as possible for input generated images. In this study, Mean Squared Error (MSE) is employed to quantify the loss, and the Discriminator's evaluations for both images are combined to derive the loss value for that specific Discriminator.

As for the configuration of the Generator's loss function, it involves placing the images generated by the Generator into the Discriminator for assessment. The goal is to ensure that the images generated by the Generator (i.e., those subjected to style transformation) closely resemble the original images from the target style dataset. In other words, it aims for the Discriminator's evaluation of generated images to be as close to 1 as possible, indicating similarity to the target style. Similarly, Mean Squared Error (MSE) is employed to quantify the loss in this context.

As the model undergoes training, the loss values progressively decrease. This can lead the Generator to prioritize generating images that confuse the Discriminator, potentially resulting in the loss of the content from the input images. However, in the context of style transfer tasks, the expectation is for the Generator to preserve the original image content while only altering the image style. To achieve this, additional loss measures are introduced to assess the effectiveness of the Generator, specifically the cycle-consistency loss [6].

The cycle-consistency loss evaluates the difference between the image transformed in terms of style and then transformed back to the original style, measuring the gap between the resulting image after two style transformations and the original image [6]. The objective for the Generator is to focus on style transformation, aiming for the generated image to closely resemble the original image after this dual transformation.

Based on the measurement of cycle-consistency loss, another set of Generator and Discriminator models is employed to perform style transformation for another style. The evaluation approach for this set of models aligns with the previous set's evaluation method. In this study, L1 loss is employed to quantify the cycle-consistency loss.

Ultimately, for assessing the loss of the Discriminator, the average value of the two Discriminators is utilized. Regarding the loss of the two Generators, due to the inclusion of cycle-consistency loss, a parameter named LAMBDA_CYCLE is introduced to balance the proportions of the two types of losses [6]. In this study, LAMBDA_CYCLE is set to 10, defining the weight of the cycle-consistency loss in relation to other losses.

During the model optimization process, Adam optimizers are employed for both the Discriminator and Generator to enhance model performance. Through iterative experimentation, the learning rate is

DOI: 10.54254/2755-2721/51/20241192

fine-tuned and eventually set at 3e-4, aiming to achieve improved training outcomes and style transfer results.

3. Result

In this chapter, graphs illustrating the loss values for the Discriminator and Generator are plotted. The iterative outputs of target images at each epoch during the training process are showcased. Additionally, the results of the test images are presented, featuring a side-by-side comparison between the original image and the generated image for visual evaluation. All experiments are conducted on Google Colab with V100 GPU, using PyTorch framework.

3.1. Performance

Figure 2 illustrates the variation in Generator loss values during the training process. The losses are computed on an epoch-by-epoch basis, with the average loss value per epoch being plotted.

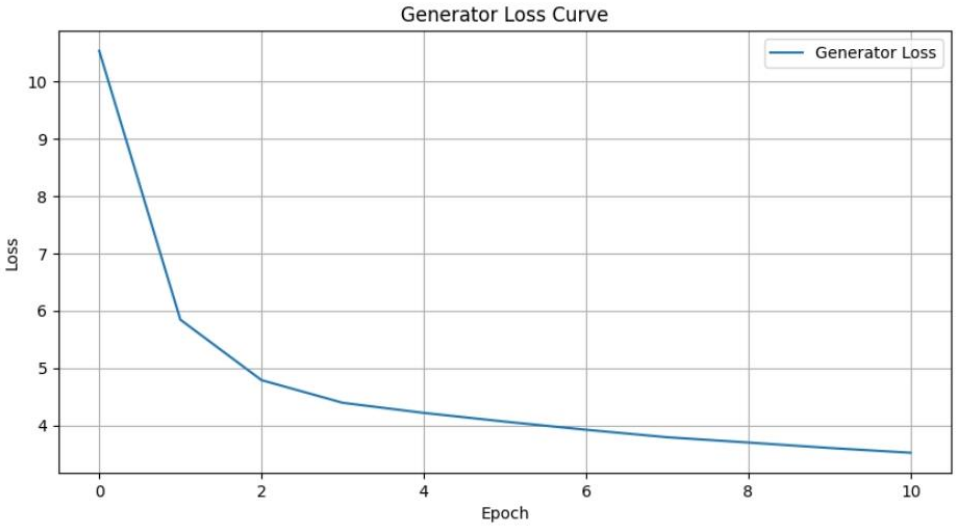

Figure 2. Loss curve of the generator (Figure Credits: Original).

Figure 3 illustrates the variation in Discriminator loss values during the training process. The losses are computed on an epoch-by-epoch basis, with the average loss value per epoch being plotted.

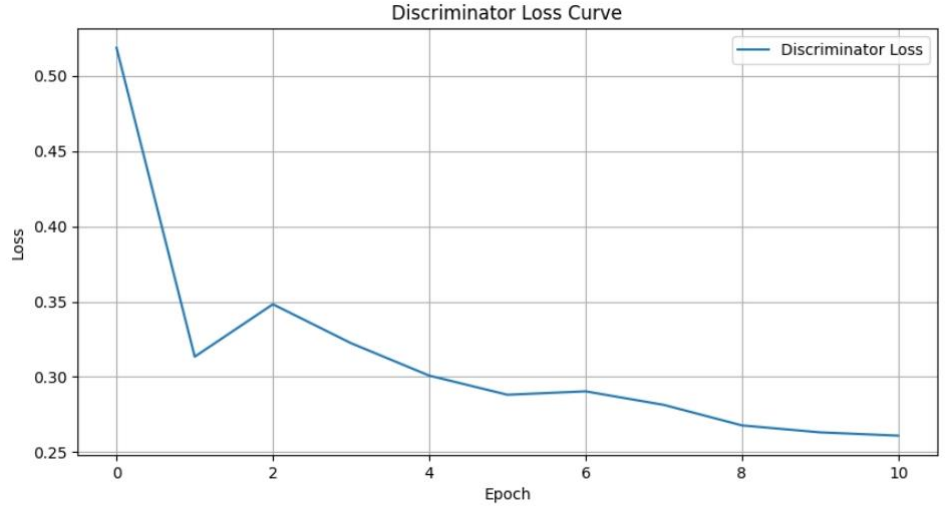

Figure 3. Loss curve of the discriminator (Figure Credits: Original).

DOI: 10.54254/2755-2721/51/20241192

The variation in model loss values reveals several insights. In the early stages of training, the Generator loss (G_loss), which assesses the effectiveness of style transfer by the generator, starts relatively high. By the 10th epoch, G_loss decreases from its initial value of 10.53 to 3.53.

On the other hand, the Discriminator loss (D_loss), which measures its ability to distinguish between original and generated images, exhibits an overall decreasing trend during training, with occasional fluctuations and temporary increases. This behavior is primarily attributed to the growing capability of the Generator to create images that confound the Discriminator, making it difficult to differentiate between original and generated images.

However, as training progresses, the Discriminator continues to enhance its capacity to recognize generated images, leading to a later decrease in D_loss. Simultaneously, the reduction in D_loss encourages the Generator to produce higher-quality and more detailed images, resulting in improved training outcomes.

3.2. Visualization

Figure 4 displays the output of style transfer results on target images at each epoch during the training process, serving as a means to monitor the effectiveness of the model's training.

Figure 4. Intermediate result of transferred images (Figure Credit: Original).

It can be observed that at epoch = 1, the image lacks discernible content, showing only basic contour information, and the ancient style exhibits primarily basic colors. At epoch = 2, the basic shape of the mountains in the image starts to emerge, and there is more diversity in the colors associated with the ancient style, although clarity remains limited. As the model continues training, both the colors and details of the mountain image undergo changes. At epoch = 7, the essential features of the mountain landscape become more apparent, and it becomes discernible that the image depicts mountains and a seascape, with the ancient style also becoming more evident. In the subsequent training process, details become increasingly clear, and the integration with elements of the ancient style improves.

Figure 5 displays the style transfer results, showcasing the model's final performance by comparing the original images with the test output images.

The generated images effectively maintain the content of the original images, ensuring that the intended content remains unchanged. The generator excels in performing style transfers, whether applied to natural landscape photos or urban street scenes. In conclusion, the generated images exhibit the artistic style of ancient Chinese artworks.

DOI: 10.54254/2755-2721/51/20241192

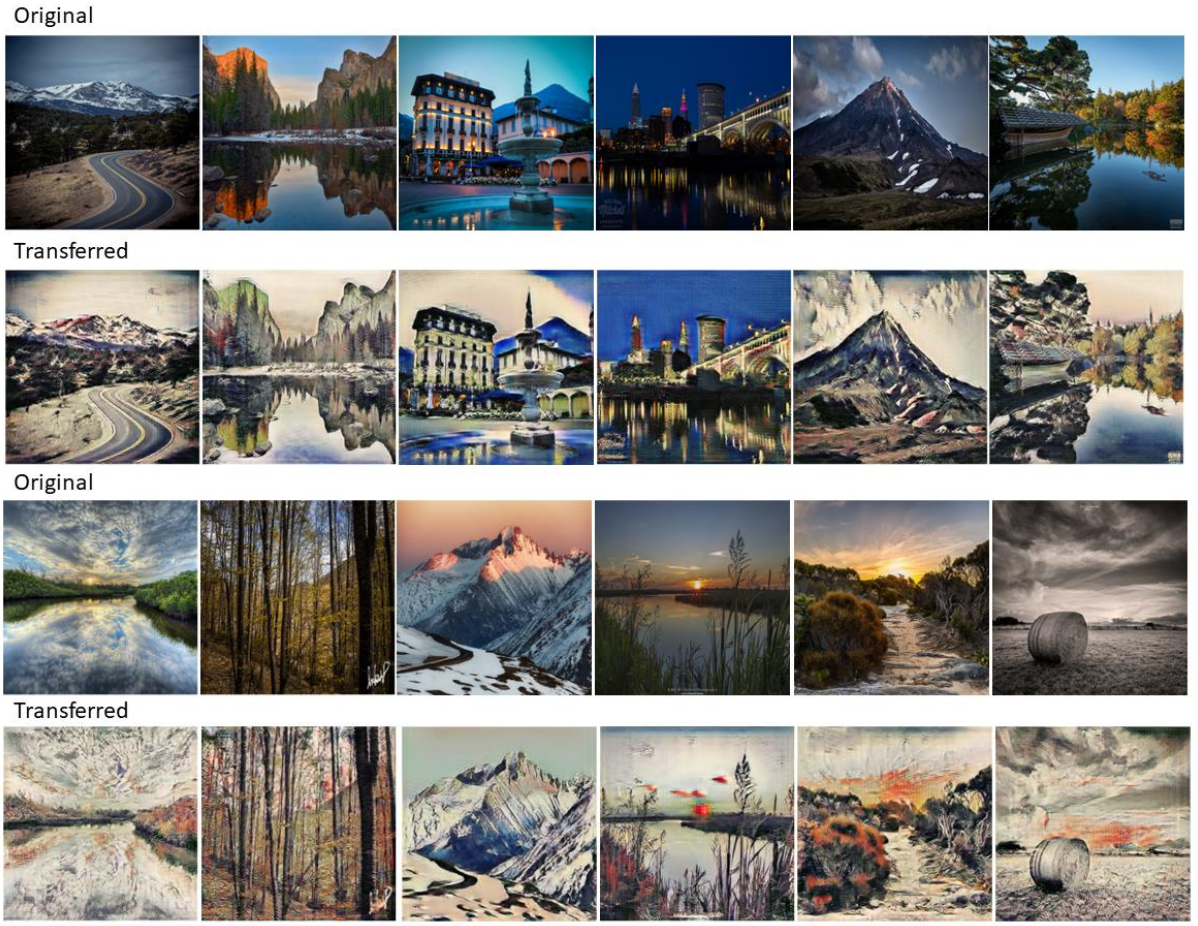

Figure 5. Visualization result (Figure Credit: Original).

4. Conclusion

Through a comprehensive analysis of intermediate model outputs, an observation of model loss curve variations, and multiple parameter adjustments, favorable model performance was achieved in this experiment. Different learning rates were experimented with, and the one resulting in the best model performance was selected. Additionally, the LAMBDA_CYCLE parameter was fine-tuned multiple times, ultimately settling on a value of 10, which produced the most desirable image results while preserving the content of the original images. During the experimental process, real-time monitoring of the model training output guided the selection of an appropriate number of epochs to stop the model training, ultimately settling on 10 epochs.

This study has demonstrated style transfer between modern photographs and Chinese ancient paintings using CycleGAN, yielding promising results. The creativity and potential applications of this approach warrant further exploration. There is room for further improvement in model performance, which can be achieved by experimenting with different loss functions and altering network architectures. Additionally, continued exploration of other style transfer tasks can enhance the model's generalization capabilities.

References

[1] Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2019). Neural style transfer: A review. IEEE transactions on visualization and computer graphics, 26(11), 3365-3385.

[2] Singh, A., Jaiswal, V., Joshi, G., Sanjeeve, A., Gite, S., & Kotecha, K. (2021). Neural style transfer: A critical review. IEEE Access, 9, 131583-131

DOI: 10.54254/2755-2721/51/20241192

[3] Cai, Q., Ma, M., Wang, C., & Li, H. (2023). Image neural style transfer: A review. Computers and Electrical Engineering, 108, 108723.
[4] Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.
[5] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1125-1134.
[6] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, 2223-2232.
[7] Sheng, J., Song, C., Wang, J., & Han, Y. (2019). Convolutional neural network style transfer towards Chinese paintings. IEEE Access, 7, 163719-163728.
[8] ukiyoe2photo, URL: https://www.kaggle.com/datasets/helloeyes/ukiyoe2photo. Last Accessed: 2023/09/22
[9] Chu, C., Zhmoginov, A., & Sandler, M. (2017). Cyclegan, a master of steganography. arXiv preprint arXiv:1712.02950.
[10] Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., & Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In International conference on machine learning, 195-204.