An image style transfer algorithm based on attention mechanism and GhostNet

Zeyu Dong[1,3,†] **and Hua Li**[2,†]

[1]Information Technology, Lowell Institute School, Coll of Professional Studies, Northeastern University, 360 Huntington Ave, Boston, MA 02115, the US
[2]Faculty of Chemical Engineering, Beijing University of Chemical technology, 15 Beisanhuan Dong Lu, Chaoyang District, 100029, Beijing, CN

[3]dong.zey@northeastern.edu
†These authors contributed equally.

Abstract. Image style transfer, which combines the content of one image with the style of another to create a new image, has many potential applications in the disciplines of image creation, creative style reproduction, animation production, and other areas. The majority of early image style transfer techniques used manual features. Due to advances in deep learning technology, the Generative Adversarial Network (GAN) method greatly increases the accuracy of image style transfer; however, when it comes to scenes with large color blocks, high resolution, or horizontal boundaries, their performance still falls short of practical application requirements. In this research, we present a GhostNet- and attention-based image style transfer technique. We specifically introduce the SELayer to learn and increase the weight value of each channel, which can improve the quality of style transfer and data processing capacity. In addition, GhostNet is included to accelerate image production even more. The findings of the experiments, both quantitative and qualitative, demonstrate that the suggested technique can enhance the impact of style transfer.

Keywords: image style transfer, GAN, attention mechanism.

1. Introduction

The goal of image style transfer is to combine the content of one picture with the style of another image to create a new image. To do this, you would combine image A's content with image B's style, discard image A's style and image B's content, and create image C. Both the traits of picture A and the substance of image B are present in image C [1]. At present, image style transfer has a good application prospect in image generation, artistic style reproduction, cartoon production and other fields, and it is a new technology with development potential.

After the historical development, there have been several different ways to realize image style transfer. In the early stage, traditional methods were adopted in style transfer, including Non-photorealistic graphics, texture transfer, CG, CV, DIP and other technologies. Most of them were not based on artificial intelligence with poor image processing effect and slow processing speed [2], which can not meet the practical application needs. Because of the technological progress of deep learning, the method based on convolutional Neural Networks greatly improves the accuracy of image style transfer.

After 2015, researchers began to use neural network technology to deal with style transfer, mainly through texture modeling-image reconstruction. This method has a wider application range than traditional methods, but the effect is not very satisfactory. One of the reasons is that most of the style transfer of this method requires paired pictures as training sets, however such data sets are hard to find [3]. After 2017, generating confrontation network (GAN) began to be widely used in the field of style transfer. Actually, generating confrontation network As a kind of neural network, GAN is actually a style transfer based on neural network. Through the game between the generator and the discriminator, the ideal style transfer pictures are finally obtained. It can achieve a good migration effect, but the model training takes a relatively long time [4].

The current style transfer methods all have a drawback, that is, the trained model can only deal with a certain kind of pictures, but the effect of other pictures with different styles is not good. The traditional GAN model has great disadvantages in dealing with the style transfer from natural pictures to children's book illustrations, while a GAN-based model: Ganilla, by integrating the ideas of CycleGAN and DualGAN, has created a new network structure, successfully solved this problem and filled the gap in this field. However, Ganilla model still has a major flaw. The author of this model mentioned in the article that Ganilla is difficult to handle black and white or solid color background images and images with clear boundaries [5]. After pre-experiments, we found that this problem is really serious, and nearly one-third of the image processing results in many data sets are very poor. If this problem is not solved, the application scope of Ganilla model will be greatly limited.

After comparing and analyzing the output results of GANILLA, it is found that the images with poor processing results often have the common characteristics of large color blocks, high degree of discrimination or horizontal boundaries. There may be two reasons for this phenomenon: First, when the original author constructs the model training model, there is not enough proportion of the images of this type of artistic style, which makes the model weak in processing this kind of scene after learning. We should find a way to obtain relatively good quality images even if some styles of images in the training set are insufficient. Second, GANILLA actually continues the conventional convolution training method, and performs conventional convolution-pooling training on the whole picture, but does not specialize some features, which will cause some excellent features to be missing due to convolution, while bad features will be enlarged due to convolution. Therefore, we expect to find a way to strengthen the cognition of features in the model. We have noticed that the attention mechanism can enhance the data processing ability by dividing the data into different channels and increasing the weight value for each channel [6]. Therefore, we plan to use SELayer module to realize the attention mechanism, so as to strengthen the cognitive ability of Ganilla model features and improve the quality of style transfer. After SELayer enhancement, the model processing effect has been significantly improved, but its image processing speed has dropped a lot. Therefore, we introduced some modules in GhostNet to reduce the calculation consumption of the model and improve the calculation speed [7].

2. Related work

CycleGAN and DualGAN are two excellent improved GAN networks. In the training process of Ganilla model adopted in this study, the ideas of these two GAN networks are adopted.

2.1. CycleGAN

CycleGAN is mainly used to solve the problem of Domain Adaptation. Before CycleGAN, related models of domain migration, such as Pix2Pix, had to use two sets of related and strictly corresponding image sets as training sets. It is difficult to find such a data set, and the appearance of CycleGAN solves this problem. The core idea of Cycle is shown in the following figure:

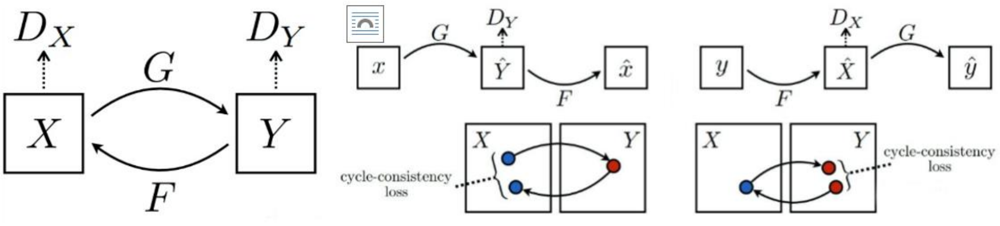

Figure 1. Schematic diagram of the principle of CycleGAN model.

The graphic depicts a ring neural network called Gan. The generator F is used to recreate the original picture input in the X domain after the generator G has generated the image in the X domain to the image in the Y domain. The generator G reconstructs the original picture input in the Y domain after the generator F generates the image in Y domain for the image in X domain. The created picture may be identified by the D x and D y discriminators to assure the image style transfer effect [8].

The loss function of GAN consists of two parts, namely:

$$Loss = Loss_{GAN} + Loss_{cycle}$$

Among them, LossGAN ensures that the generator and the discriminator evolve with each other, thus ensuring that the generator can produce more realistic pictures; Losscycle guarantees that the output picture of the generator is different from the input picture only in style, but the content is the same. This is the basic idea of cycle-consistency loss [8].

2.2. DualGAN

Enhancing the fundamental GAN network is the DualGAN algorithm. A complicated network with two generators and two discriminators is created by the algorithm by combining two Gans. As an illustration, consider images X and Y. Image Y is the genuine photo, whereas image X is the sketch. The initial generator GA converts X into Y, and the discriminator DA is in charge of differentiating the newly changed image Y. Additionally, there is a collection of generator-discriminator combinations that are applied to Y to X conversion and image quality authentication[9]. DualGAN has solved the common problems in GAN that it is difficult to train and can't optimize [10], and provided a new idea for "translation" between pictures. Unsupervised learning algorithms such as DualGAN may greatly reduce the cost of tagging, and it will have more important application value if its stability can be further improved.

3. Method

3.1. Baseline method of Ganilla

In this paper, we improve the image effect and image processing speed based on Ganilla model [5], which is the most representative image style transfer algorithm. Ganilla is a new image transmission model. In the problem of image-to-illustration, unpaired image-to-image translation models, such as CycleGan [3], DualGan [4], CartoonGan[5], etc., all have a problem, that is, they can only satisfactorily complete one of the tasks of style transfer or image transfer. Therefore, Ganilla's original design purpose is to complete the style transfer from nature pictures to illustrations. This model can not only complete the style transfer task well, but also retain the content of the original pictures. In Ganilla model, a novel generator network is proposed to achieve a good balance between style transfer and content retention. In addition, Ganilla puts forward a new framework for evaluating image generation model, which can evaluate the generated results based on both content and style [5].~~5~~

Figure 2 depicts the network structure of Ganilla. Ganilla adopts low-level characteristics so that content is maintained while styles are changed. The downsampling stage and the upsampling stage are the two components of the model, as shown in Figure 2. The modified ResNet-18 network is used for the downsampling stage, and each layer is connected to the layer before it in order to preserve the image's morphological details, edges, forms, and other information. A 7x7 convolution layer, instance normalization, ReLU, and max pooling layers are all included in the downsampling step before moving on to four network layers, each of which contains two residual blocks. Convolution is the first layer in each residual block, which is then followed by instance normalization and ReLU layer. The output is combined with the input of the residual block after the convolution layer and instance normalization. The ReLU layer then receives the final convolution of the spliced tensor. The author uses a technique that uses low-level features to improve the image content retention ability in the up-sampling stage where he connects the output of each network layer from the down-sampling stage to the summation layers. Four connected convolution, upsampling, and summing layers make up the upsampling stage. In every stage of the upsampling process, the convolution filter kernel is 11. The transformed three-channel image is then output using a convolution layer with a 77 core. A 7070 PatchGAN with three convolution blocks that each have two convolution layers makes up the discriminator network. The initial convolution block's filter size is 64, and successive blocks' filters grow in size by a factor of two. The GANILLA model is trained using the concept of cycle consistency. While the second group F uses the input image as the target domain and attempts to construct the source image in a circular manner, the first group G tries to conceal the source image within the target domain.

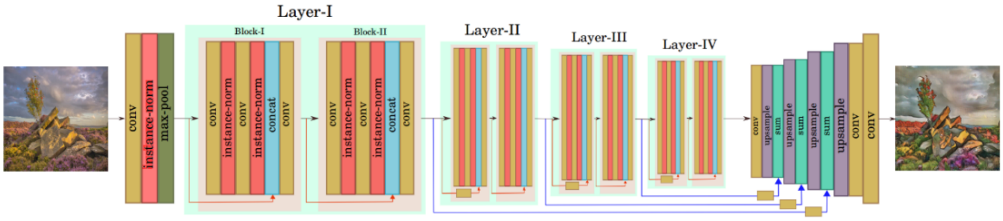

Figure 2. Ganilla's network structure.

3.2. *Model optimization*

3.2.1. Attention model. For CNN network, the core calculation is convolution operator, which learns from the input feature graph to the new feature graph through convolution kernel. In essence, convolution is the feature fusion of a local area, which includes feature fusion in space (H and W dimensions) and between channels (C dimension).

Improving the receptive field for convolution operations entails fusing more spatial elements or extracting multi-scale spatial information. Convolution by default merges all of the input feature graph's channels for feature fusion of channel dimension. In the MobileNet network group channel, group convolution and depthwise separable convolution are mostly used to lighten the model and reduce processing. In the hopes that the model would automatically recognize the significance of various channel attributes, SENet network pays close attention to the relationships between channels. This issue can be resolved using the Squeeze-and-Exclusion (SE) module suggested by SENet [11], as depicted in the following picture.

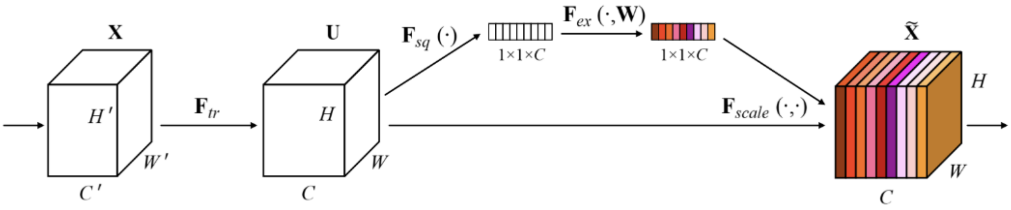

Figure 3. Schematic of the squeeze-and-Exclusion (SE) module.

In this paper, we have adopted SE module many times. The module is imported into the original Ganilla network, and the output of each Layer is used as the input content of the SE module, and the output of the SE module is introduced into the convolution Layer of the next layer or the up-sampling stage. In this study, the attention mechanism is realized through this module, and the weight is added to each channel according to the difference in the importance of features among channels, thus realizing the enhancement of low-dimensional features in high dimensions, improving the performance of excellent features to a certain extent, while weakening the appearance of bad features.

3.2.2. GhostNet. After a picture is extracted by neural network, many feature maps can be obtained. There will be some similarities in the feature graphs, which is the jumble of feature graphs in neural networks. It is possible to perform simple Cheap Operations on one of the feature graphs to generate more similar feature graphs, so that more feature graphs can be generated with fewer parameters, and the similar feature graphs can be regarded as each other's ghosts, which is the basic logic of GhostNet. Convolution as we know it is replaced by the Ghost Module. Prior to using depth separable convolution (layer by layer convolution) to obtain additional feature maps, the common convolution is used to reduce the number of channels in the input image. Then, several feature maps are concatenated to create a new output [7].

In this paper, we introduce the Ghost Module module in GhostNet, and replace the conventional 3x3 convolution layer in the original model with Ghost Module. Its parameters are set as follows: kernel_ size = 3, ratio = 2, dw_ size = 3, stride = 1, relu = true, padding = 0. By adding GhostNet, the total amount of data to be processed is reduced, and to some extent, the problem of slow operation speed of the model is improved.

4. Result

4.1. Summary of results

In this paper, by improving the original Ganilla network, the style transfer effect of Ganilla for images with large color blocks and high difference is improved. For the following image, we can find that the original Ganilla method will produce a lot of meaningless noise and color difference when transferring styles, and the whole picture will appear unnatural green. After introducing the attention mechanism of SENet, it can be clearly seen that the content of the original image has been preserved more. The improved result is clearer overall, and the overall tone is closer to the original image.

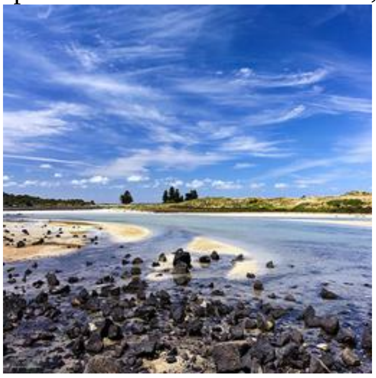

Figure 4. Original image.

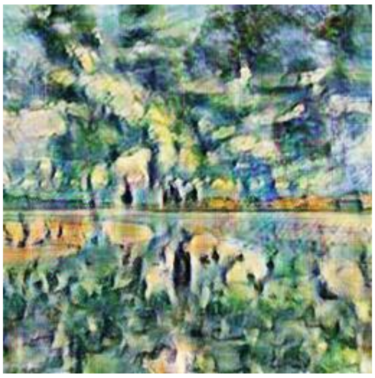

Figure 5. The style migration effect of the model before improvement.

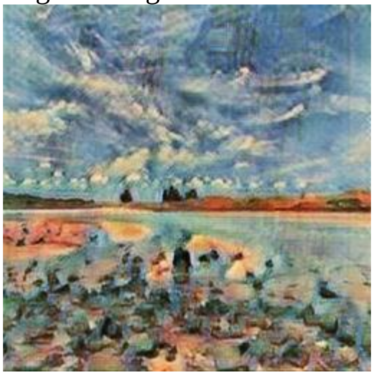

Figure 6. The style migration effect of the model after improvement.

In order to quantitatively analyze the improved effect, this paper uses Beyond Compare4 [12] to compare the generated image with the original image. By overlapping the images, the software will mark the differences between the two images in red after setting the tolerance. In this way, we can judge whether the retention degree of image information in the improved model has been improved. The results are as follows:

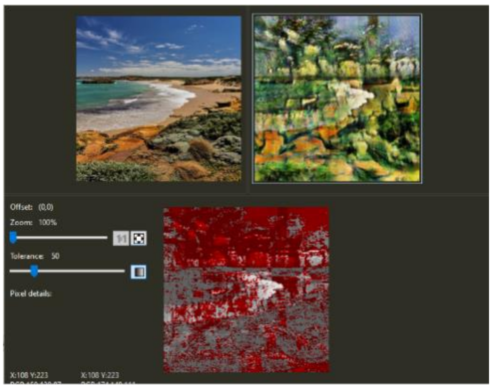
Offset: (0,0)
Zoom: 100%
Tolerance: 50
Pixel details:
X:108 Y:223 X:108 Y:223

Figure 7. Analysis results of the model before improvement.

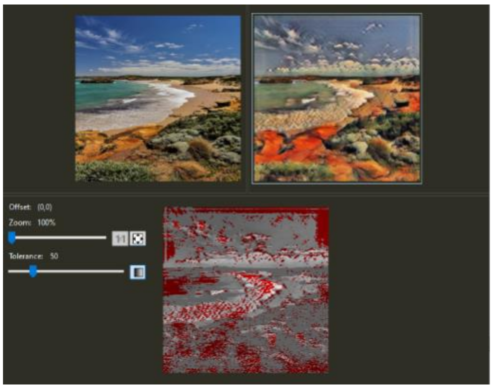
Offset: (0,0)
Zoom: 100%
Tolerance: 50

Figure 8. Analysis results of the model after improvement.

When the tolerance is also set at 50, the original Ganilla model has a large number of red areas in the sky, which indicates that this area is quite different from the original image, and the content of the original image is not well preserved. In the improved output results, the difference in this area is obviously reduced, and there are still obvious differences only in the border. However, the whole content of the image is still fully preserved, so we think that the introduction of SENet significantly reduces the loss of the original Ganilla model to the image content of similar large-color images.

In terms of style transfer effect, we use geotests.net to analyze the color of HSL image. The principle is to divide the image into small color blocks one by one, then output the HSL value of the color block, and list the distribution of the color blocks in the coordinate system. By analyzing the color composition of the image, we can clearly see the effect of image conversion according to the specified style.

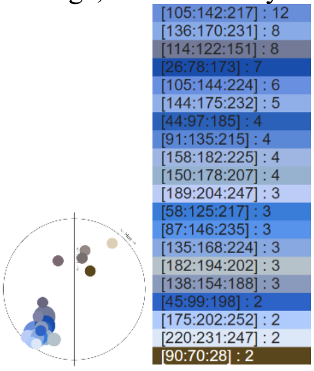

Figure 9. Analysis results of original image.

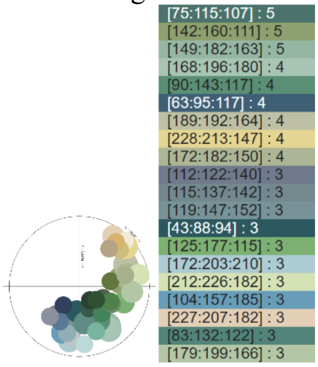

Figure 10. Analysis results of the model before improvement.

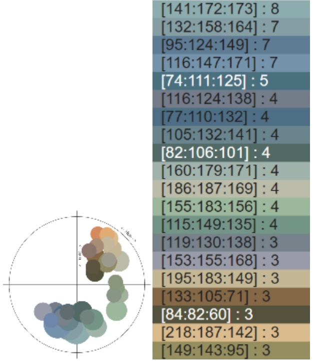

Figure 11. Analysis results of the model after improvement.

In the figure above, the contents of the "[]" symbol represent the HSL value of the color, and the numbers in the figure below represent the number of blocks of the color.

We found that compared with the original model, the improved Ganilla is more similar to the original image in color composition and distribution, but it still changes the overall color tone. At the same time, it is similar to the original Ganilla. We think it shows that the output result of the new model has successfully transferred the style of the image, and its conversion degree is slightly lower than that of the original Ganilla, rather than just adding a filter effect to the original image.

At the same time, the optimization of the network model brought by GhostNet has been improved compared with the original Ganilla. In the process of model training, we used a RTX2080Ti graphics card with single floating-point performance of 13.4 TFLOPs as a test platform. In the experimental process, the operation speed of the improved model is increased by 2% compared with that of the original Ganilla model. Because we also introduced the SENet module in the testing process, the data processing capacity is already higher than that of the original Ganilla model, so GhostNet should improve the original model by more than 2%.

4.2. Discussion

Although this method improves the retention of the original image content to a certain extent, when faced with this kind of image, obvious bright lines will appear at the edges of large color blocks and images. There is still much room for improvement in the preservation of the original content. In addition, there will be a small probability of generating square noise in the image, and a satisfactory solution can not be found.

5. Conclusion

In this paper, we improve the problems faced by Ganilla when dealing with images with large color blocks, high degree of discrimination, or horizontal boundaries. Ganilla adopts the conventional convolution layer training method, which makes the image lose more content in the convolution process. To overcome this problem, we introduced SENet into the original model, added attention mechanism to Ganilla, and realized feature enhancement from low dimension to high dimension. To some extent, the retention of image content is improved.

Another problem of Ganilla is the increase of computation due to complex network structure. In order to solve this problem, we introduce the GhostNet module, and replace the traditional convolution, and use linear operation to generate more similar feature maps from one feature map. Finally, the computing speed is improved by about 2%.

References

[1] Mou Jinjuan. Research on image style transfer technology based on deep learning [J]. Electronic Components and Information Technology, 2019 (04): 82-85. doi: 10.19772/j.cnki.2096-4455.2019.4.024.

[2] Jetchev N , Bergmann U , Yildirim G . Copy the Old or Paint Anew? An Adversarial Framework for (non-) Parametric Image Stylization[J]. 2018.

[3] Gatys L A , Ecker A S , Bethge M . Image Style Transfer Using Convolutional Neural Networks[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[4] Goodfellow I , Pouget-Abadie J , Mirza M , et al. Generative Adversarial Nets[C]// Neural Information Processing Systems. MIT Press, 2014.

[5] Hicsonmez S , Samet N , Akbas E , et al. GANILLA: Generative Adversarial Networks for Image to Illustration Translation[J]. Image and Vision Computing, 2020.

[6] Jaderberg M , Simonyan K , Zisserman A , et al. Spatial Transformer Networks[C]// MIT Press. MIT Press, 2015.

[7] Han K , Wang Y , Tian Q , et al. GhostNet: More Features From Cheap Operations[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

[9] Zili Yi, Hao Zhang, Ping Tan, Minglun Gong. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation

[10] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, Wei-Ying Ma. Dual Learning for Machine Translation

[11] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[12] Beyond Compare 4 Current Version: 4.4.3, build 26655, released July 20, 2022 Scooter Software, Inc. 625 N Segoe Rd, Suite 104 Madison, WI 53705 USA