Image Style Transfer based on Generative Adversarial Network

Chan Hu[1] ,Youdong Ding[1] ,Yuhang Li[1]
1. Department of film and television engineering, Shanghai University
1217136173@qq.com

***Abstract*—The traditional style transfer based on GAN model is limited by paired images. CycleGAN solves this problem effectively, but its structure is complex and training time-consuming. This paper proposed a style transfer model based on generative adversarial network, which abandons the redundant structure of two GAN models trained by CycleGAN, and trains only one generator and one discriminator without pairing image samples. And the semantic content between the input image and the generated image is constrained by the VGG network feature map. In order to accelerate the convergence of the model, a pretraining stage is introduced.The experimental results show that the proposed model is effective than CycleGAN,and outperforms state-of-the-art methods.**

Keywords—style transfer; CycleGAN; VGG network

I. INTRODUCTION

As a major field of computer graphics, Non-photorealistic rendering (NPR) has been devoted to the simulation of different styles. For example, the simulation of ink painting, oil painting and watercolor painting. These technologies are also very mature, and the achievements are also applied to all walks of life. Ink painting, oil painting and other artistic works have obvious style and characteristics, but the imitation of works of specific artists needs a lot of research. In recent years, with the rise of deep learning and the enhancement of computer hardware capabilities, some image style transfer methods based on learning have been proposed. The original neural style transfer can only transfer the specified style to the specified content image, which has great limitations. Later, real-time style transfer based on loss perception can quickly transfer the specified style to any content image compared with neural style transfer. It has a great improvement, but it is only limited to the transfer of specific style, unable to achieve multi-style transfer. Then, the meta network effectively solves the problem of multi-style transfer, which can quickly transfer any style to any content image. These methods of style transfer based on convolutional neural network can only learn the style of a single style image, but not the overall painting style of a artist There are still many problems in extracting the overall style of the artist from many paintings. In recent years, the Generative Adversarial Networks (GAN) is widely used in the field of image processing. Compared with other generation models, it gradually pushes the training image closer to the data distribution of the target image in a semi supervised or unsupervised way. However, the original GAN model can only generate images randomly, and need a pair of data sets that is obviously difficult to obtain. CycleGAN effectively resolves the constraint of paired data sets and generates high-quality style transfer results in a cyclic manner by training unpaired photo collections and selected artist's painting. However, CycleGAN has a complex structure and needs to train two pairs of generator - discriminator models. Therefore the training time is long and the convergence is slow. In this paper, the cycle architecture of CycleGAN is abandoned, because the backward mapping from style image to input image is unnecessary in style transfer. This method uses perceptual loss to constrain the semantic content between the input image and the output image, and uses the adversarial loss of GAN model to push the data distribution of the input image to the data distribution of the style image, so as to realize the style migration. At the same time, a pretraing stage is introduced to accelerate the convergence of the model. Compared with the state-of-the-art methods, this method achieves better style transfer results, less training time and faster model convergence.

II. RELATED WORK

Image style transfer refers to mapping the style of one image to another content image, and ensuring that the semantic content of the content image does not change. Style transfer is actually the extension and expansion of Non-photorealistic rendering [1~2]. Non-photorealistic rendering, also known as style rendering, is proposed relative to photorealistic rendering. It refers to the computer-generated image technology that does not have photo like reality, but has a specified style. Its goal is not to generate whether the image is realistic, but mainly to simulate the image art style. Traditional methods develop special algorithms for specific styles, which are time-consuming and laborious, and are not easy to expand.

With the rise of deep learning technology, some data-driven style transformation algorithms have been proposed. For example, gatys et al.[3] first proposed the neural style transfer method (NST), which uses the pretrained VGG[4,5] network to separate and extract the content and style of the image, and generates the image by fusing different styles and contents. This method can automatically transfer all kinds of art style images, but it needs the specified style image and the specified content image, and the two images should be reasonably similar. In addition, each generated image has to go through multiple iterations, which is very time-consuming. Johnson et al.[6] used perceptual loss to train a feed-forward network to realize real-time image style transfer, which solved the time-consuming problem of gatys et al. method. However, once the network training is completed, only the transfer of specified style can be realized. If other styles are to be realized, additional training network is needed. In order to realize the transfer of any style, Shen et al.[7] proposed a new method of specifying network parameters through Meta network to realize fast style transfer of any style and any content.

978-1-7281-4390-3/20/$31.00 ©2020 IEEE

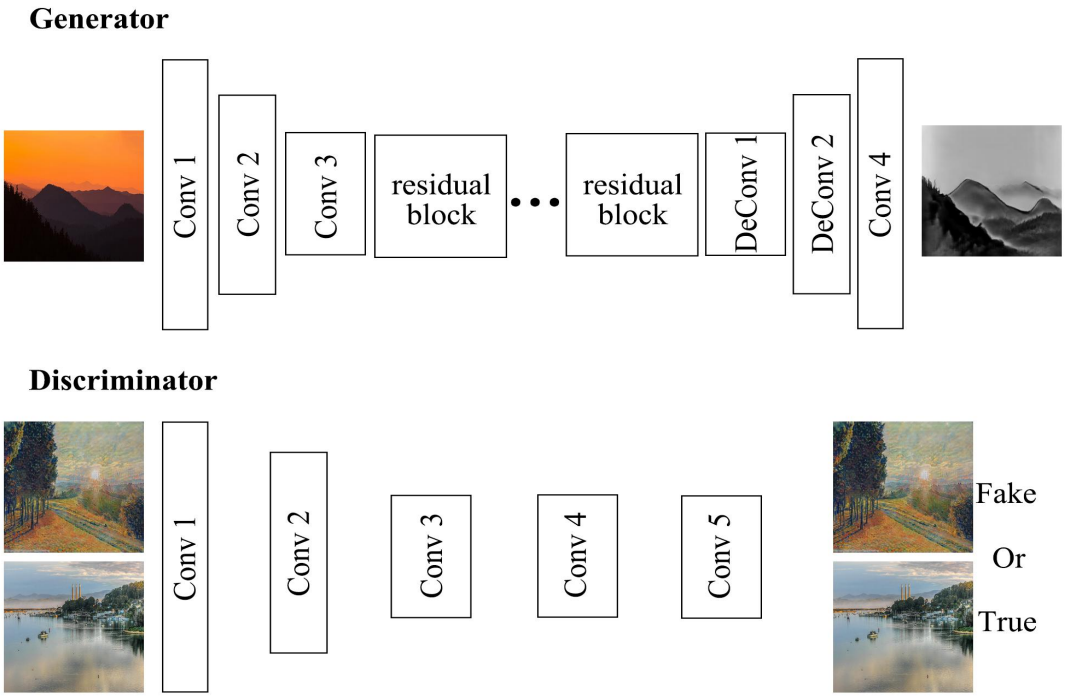


Fig. 1. The architectures of Generator and Discriminator.

Goodfellow et al.[8] proposed the new model of GAN, which consists of two networks: generator *G* and discriminator *D,* whereby the adversarial loss provided by the discriminator pushes the generated images towards the target manifold. Radford et al.[9] proposed the deep convolutional Generative Adversarial Networks (DCGAN), which combines GAN with the deep convolutional network structure for specialized image generation. No matter GAN or DCGAN, they can only generate random data, but can't generate specific data exactly. Therefore, Mirza et al.[10] proposed conditional GAN (CGAN) with conditional constraints, so that the model can produce results that meet the conditions; Chen et al.[11] proposed CartoonGAN to realize the cartoon style of photos by designing an edge promoting loss. Isola et al.[12] proposed image to image translation tool based on CGAN, but it needs paired data in the training process. Usually, paired data is very difficult to obtain. therefore, Zhu et al.[13] proposed CycleGAN, which does not require paired images and can also realize the transformation of one kind of images to another.

III. Proposed Method

A. Network Architectures

The GAN model in this paper contains two CNNS, one is generator *G*, which is used to learn the mapping between the image and the specified style image. The other is discriminator *D*, which is used to judge whether the image is generated by *G* or the real style image. The generator and discriminator structures are shown in Fig. 1.

The generator network *G* can realize the transformation of input image to style image. An image with a size of 256*256 and a channel number of 3 input generator network. First,image passes through a convolutional layer and then through two down-convolution blocks to obtain 256 feature vectors of 64×64. At this point, the image resolution gradually decreases and the number of channels increases. Afterwards, after 6 residual blocks[14], the extracted feature vectors are fused. By the struct of "shortcut connection", the residual block attach the features of the previous layer into the current layer, which can effectively relieve vanishing gradient[15,16] issue caused by network deepening. Finally, the output style image is reconstructed by two up-convolution block.Instance normalization[17] and Leaky ReLu (LReLU)[18]is widely used in training deep CNNs.We integrate these techniques in our method.The specific parameters of the generator are shown in table 1

TABLE I. The specific parameters of the generator

layer	activation size
input	3×256×256
64×7×7conv,stride1	64×256×256
128×3×3conv,stride2	128×128×128
256×3×3conv,stride2	256×64×64
6×Residual block	256×64×64
128×3×3deconv,stride2	128×128×128
64×3×3deconv,stride2	64×256×256
3×7×7conv,stride1	3×256×256

The discriminator network *D* is used to judge the probability that the image generated by the generator is a real style image. It is a simple binary classification model and its structure is a convolutional neural network. The input image passes through some convolutional layers to obtain the classification response.

Fig. 2. The experimental results

B. Loss Function

The loss function of this model is in Eq.(1), which is composed of two parts: the adversarial loss L_{adv} and the content loss L_{con}. The adversarial loss put G from the input images towards the target manifold. The content loss is used to constrain semantic content between the input images and output images. λ is a hyper parameter. It is used to control how much input image content is retained in the image stylization. Its value varies with the specific style transfer. If you want to keep a more style of style image, you can reduce its value, and if you want to retain more content of the input image, you can increase its value.

$$L(G,D) = L_{adv}(G,D) + \lambda L_{con}(G,D) \tag{1}$$

The adversarial loss is the core loss in GAN network. Its value reflects what extent the image generated by the generator looks like a style image. The task of the discriminator D is to try to distinguish whether the input image is generated from generator or from true style image. In the process of rivalry game, the generator keeps learning to produce images, and the discriminator keeps improving its ability to identify images, and the two will eventually reach a balance. The adversarial loss in Eq.(2)

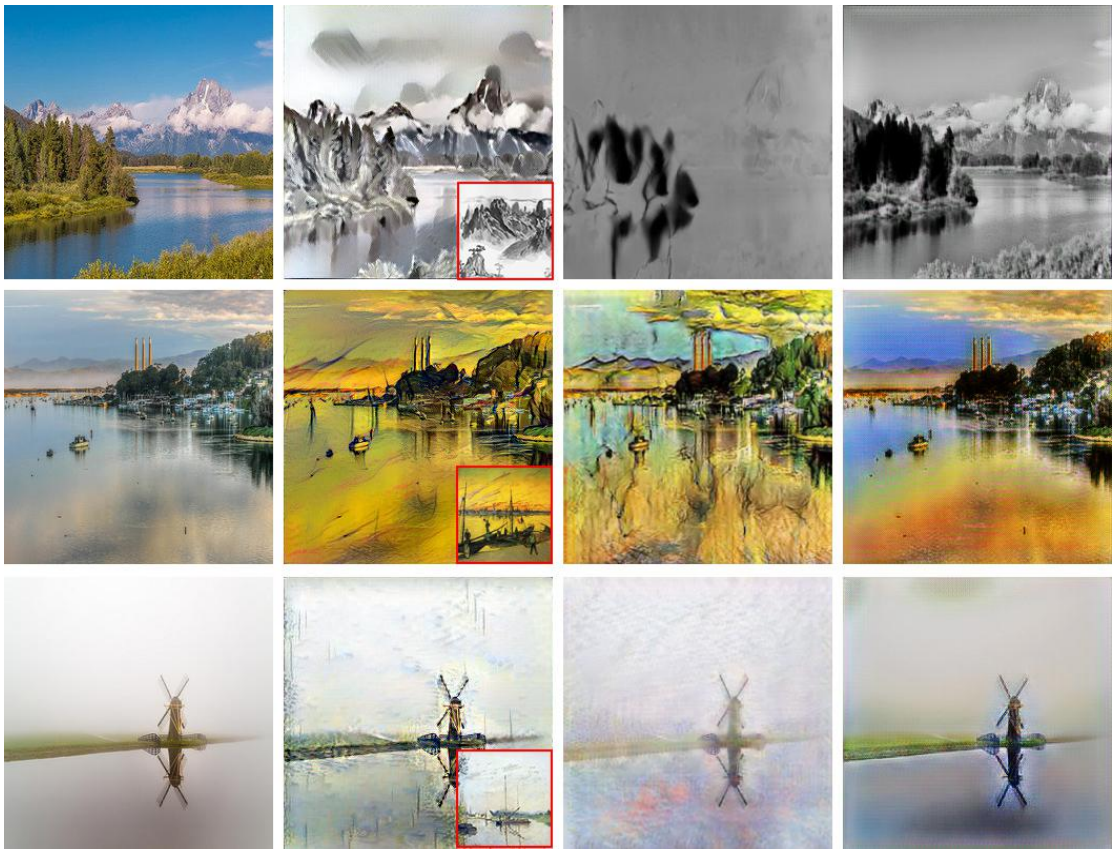
$$\begin{aligned} \min_G \max_D V(D,G) = \\ E_{y_j \sim P_{data}(y)}[\log D(y_j)] + \\ E_{x_i \sim P_{data}(x)}[\ln(1 - D(G(x_i)))] \end{aligned} \tag{2}$$

where x refers to a picture of the real world. $P_{data}(x)$ represents the distribution of photo data. y represents the image of the specified style. $P_{data}(y)$ represents the distribution of style images. E is the Expectation. ~ indicates obedience. For G, it is a minimum of the above equation, and for D, it is a maximum of the above equation. Through the rivalry game, G learns the distribution of the style image. The above equation is the adversarial loss function of GAN

In order to achieve the correct transformation between the real photo domain and the style image domain, an important principle is to ensure that the generated style image maintains most of the semantic content of the input image. In order to avoid too much content loss in the generated image after image style transfer. This paper introduces a content loss function in Eq.(3)

$$L_{con}(G,D) = E_{x_i \sim p_{data}(x)}[\| \phi(G(x_i)) - \phi(x_i) \|_1] \tag{3}$$

Where Φ refers to the feature extraction network. This paper uses the trained VGG network.l refers to the feature maps of a specific VGG layer. The high level feature map of



(a)input photo (b)NST (c)CycleGAN (d)our method

Fig. 3. Style transfer results of different methods

VGG is used to constrain the image semantic content between input images and output images.

C. Pre-training Phase

Gan model is highly nonlinear, and all parameters are randomly initialized, so it is easy to trapped at the local minimum in the optimization process. In order to accelerate the model convergence, a pre-training phase is introduced. Since the goal of generator network *G* is to reconstruct the input image with the specified style, the semantic content does not change much. Therefore, only the content loss is used to reconstruct the image in the pre-training phase. 10 iterations in the pre-training phase are enough to generate images with reasonably similar content. Similar work can be seen [3,11],which uses the content image to initialize the result image to improve style transfer quality. This pre-training phase can help the model quickly converge to the style image domain.

IV. EXPERIMENT

The experiment was implemented in pytorch framework and python language. Experimental environment:Win10 operating system,NVIDIA GTX 1070 Graphics card.

A. Data

The training data used in the experiment included real-world photos,ink paintings, Van Gao paintings and Monet paintings. The test data only includes real-world photos.Each image is 256×256 in size.

There were 3,750 photos downloaded from Flickr, 3,000 for training and 750 for testing.The ink paintings data set contains 2000 ink paintings, which are taken from the key frames of the ink wash animation "*little tadpole looking for his mother*" etc. The extracted key frames are grayed. 400 Van Gao paintings and 1072 Monet paintings came from the work [13]. In the experiment, the proposed method was used to realize three styles transfer: photo to ink painting, photo to Van Gao, and photo to Monet. In the training process, no matching relationship is specified between different types of images. The experimental results are shown in the Fig.2

In the Fig.2, the first column is the input image, the second column is the transfer result of ink painting style, the third column is the transfer result of Van Gao style, and the fourth column is the transfer result of Monet style. All generated images are well preserved from the input image.

B. Comparison with other methods

We evaluated the effectiveness of our method by comparing it with other style transfer methods, namely NST [3]and CycleGAN[13]. We show example style transfer results of different methods in Fig.3. The first column represents the input image. The second column is the result of NST method. The third column is the result of CycleGAN. The fourth column is the our method experimental results. The first row is the transfer of ink painting style. The second row is the transfer of Van Gao style. and The third row is the transfer of Monet style.

Since NST can only learn the style of a single image and transfer the style to the content image, the small graph in the bottom right corner of the second column of the result image represents the style image. In order to make a fair comparison, the selection principle of the style image is to try to choose the image that is reasonably similar to the content image. This is the classical method of style transfer. But it can only learn the style of a given picture, not all the styles of an artist's work, and it takes a long time to generate a picture. The third column is the experimental results of CycleGAN. It can be seen that CycleGAN can learn the style of style images very well. but a lot of content is lost. many generated results cannot be recognized without giving the input image. The fourth column is the result of our method, which preserves the semantic content of the input image and learns the specified style.

PSNR and SSIM are used to calculate the style transfer results of our method and other methods. Table 2 shows the experimental evaluations by PSNR and SSIM.

TABLE II. EXPERIMENTAL EVALUATIONS BY PSNR AND SSIM

style	criterion	NST	CycleGAN	Our method
ink painting	***PSNR***	10.0411	14.5037	15.8849
	SSIM	0.4729	0.5671	0.7416
Van Gao	***PSNR***	12.2032	15.3002	15.9736
	SSIM	0.5121	0.5392	0.5938
Monet	***PSNR***	14.3112	16.0812	14.4027
	SSIM	0.6526	0.6581	0.6454

As can be seen from the data in the table 2, in the Monet style transfer experiment, CycleGAN is slightly higher than our method on PSNR and SSIM, but in the ink painting and Van Gogh style transfer experiment, the PSNR and SSIM of our method are higher than the other methods.

Our method has the same characteristic of not requiring paired images as CycleGAN. However, our method takes less training time. For each iteration, the training time of our method is about 345.56s, while the training time of CycleGAN is about 795.75s. The training time of our method is about half of that of CycleGAN. This is because CycleGAN needs to train two GAN models for the bidirectional mapping, which seriously slows down the training process. In our method, the backward mapping from the style image to the input image is not required. So the cyclic architecture of CycleGAN is abandoned. The semantic content between input image and output image is constrained by the high-level feature map of VGG network. Similar work is also found in this paper [13].

V. CONCLUSION

This paper proposed a style transfer model based on GAN to realize the transformation of real world photos to the specified style images. This method eliminates CycleGAN's cyclic architecture, greatly reduces the training time. Our method is not restricted by paired image samples. In order to accelerate the convergence of the model, the pre-training stage is introduced. our model can learn the style features of the style image better and more effectively. The experimental results show that this method can better transfer the style to the new image.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61402278, 61303093).

REFERENCES

[1] GOOCH B, GOOCH A. Non-photorealistic rendering[M]. AK Peters/CRC Press, 2001.

[2] STROTHOTTE T, SCHLECHTWEG S. Non-photorealistic computer graphics: modeling,rendering, and animation[M]. Morgan Kaufmann, 2002.

[3] Gatys, A. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423, 2016.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR,abs/1409.1556, 2014.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh,S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein,et al. Imagenet large scale visual recognition challenge.International Journal of Computer Vision, 115(3):211-252,2015.

[6] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, pages 694 – 711, 2016.

[7] Falong Shen, Shuicheng Yan, Gang Zeng. Meta Networks for Neural Style Transfer. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems 27, pages 2672–2680. 2014.

[9] Alec Radford, Luke Metz, Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[10] Mehdi Mirza, Simon Osindero. Conditional Generative Adversarial Nets. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[11] Chen Y , Lai Y K , Liu Y J . CartoonGAN: Generative Adversarial Networks for Photo Cartoonization[C] In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired imageto-image translation using cycle-consistent adversarial networks. In International Conference on Computer Vision,2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770-778, 2016.

[15] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa. Learning to simplify: Fully convolutional neural networks for rough sketch cleanup. ACM Transactions on Graphics,35(4):121, 2016.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich.Going deeper with convolutions. In IEEE Conference onComputer Vision and Pattern Recognition (CVPR), pages 1-9, 2015.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International Conference on Machine Learning, pages 448-456, 2015.

[18] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlin earities improve neural network acoustic models. In Interna tional Conference on Machine Learning, volume 30, 2013.