LWComicGAN: A Lightweight Method for Realizing Scene Animation

Qingyu Yang ,Wei Chen ,Yichao Cai ,XinYing Liu ,Taian Liu ,Ge Wang*
College of Intelligent Equipment, Shandong University of Science and Technology, Tai'an, China
1016685691@qq.com, 275219349@qq.com, 565135882@qq.com, xinying0@aliyun.com, LTA999@163.com
* Corresponding author: Ge Wang Email:wanggeg@163.com

2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) | 978-1-6654-2207-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITAIC54216.2022.9836461

***Abstract*—The style transfer algorithm was originally proposed to solve the generation problem of art paintings. In recent years, the generation of animation style images has gradually become a hot research direction. The content presented in many animation film and television works is fascinating. In order to satisfy people's desire to turn real scenes into animation scenes and reduce the workload of animation producers, a Light Weight Animation Generated Adversarial Network (LWComicGAN) is proposed, which can reduce the amount of parameters and enable low-memory devices to complete network training. An optional instance layer normalization function is designed to adapt the input of each layer, and an optional instance layer residual block is proposed. The LWComicGAN algorithm uses the objective function of WGAN-GP and other various loss functions as the total loss, and also considers the gradient penalty mechanism in discriminator. The former guarantees the generation quality of all aspects of the image, and the latter guarantees the stability of the training process. The effectiveness of the proposed algorithm is verified after animation transfer experiments of realistic landscapes and characters, and we have produced an ink painting dataset and completed ink animation style transfer.**

Keywords—deep learning; image style transfer; generative adversarial networks; photo animation; LWComicGAN

I. INTRODUCTION

In recent years, the neural style transfer task is developing continuously. In the initial stage, the image translation task needs to be modeled manually according to the content of the image, which consumes a lot of manpower and material resources. After the development of deep learning, style transfer task has been improved qualitatively.

Convolutional neural networks can perform artistic style transfer of images and videos [1~2] , but the generated images are often uncertain, and the image quality cannot be guaranteed and is time-consuming. Johnson et al. [3] proposed a method for real-time style transfer, but the generated images have ambiguous semantics and missing content. Cui et al. [4] proposed a bilateral convolutional block that preserves the full semantics in the target image, but the image content edges are not smooth. Castillo et al. [5] first segmented the converted style image, and then used Markov random occasions to merge the image into the original image, and can achieve a smooth transition effect on the edge.

Generative Adversarial Network (GAN) [6] plays a pivotal role in image transfer. The method of Chen [7] et al. can obtain relatively good synthetic animation images, but the training time is too long. Pix2Pix [8] can perform style transfer on images with specific contours, but it is limited to one-to-one form, and data acquisition becomes a difficult problem. The dataset of CycleGAN [9] doesn't have pairwise relationships, but is limited to the transformation between the same images, which usually does not work well for the transformation between different objects.

In order to make the image animation effect better, this paper proposes a lightweight network architecture LWComicGAN and designs an optional normalization function to determine which normalization is suitable for the input tensors in each convolutional layer, and apply the function to the convolutional layers and residual blocks. In this paper, various loss functions are used in the discriminator of the network to reduce the difference of the images, and the gradient penalty mechanism is used to prevent the generation of gradient explosion. We use frame interception to create a dataset for Chinese ink animation movies, and implement ink animation images on LWComicGAN.

II. RELATED WORK

The generative adversarial network is based on a dynamic equilibrium. After years of development, GAN has made many excellent results in the field of animation con-version. Zhao et al. [10] proposed a loss function combining local and global styles to obtain structural information of cartoon images. Cui et al. [11] achieved the migration of anime line coloring for clothing, hair color, etc. by colorizing the line content of the target image. Gong et al. [12] used spatial transformers and attention networks to create exaggerated styles with distorted faces. Wei et al. [13] used some densely connected network layers in DenseNET to deepen the network and improve the quality of image generation.

In addition to the above-mentioned research on the generation of anime characters, there are also some studies on the animation of landscape photos. Wang et al. [14] separated three white-box representations from

images for cartoonization transfer. Ye et al. [15] used the construction of multiple decoders and a shared encoder to generate cartoonizations of various photos. The algorithm of Chang et al. [16] can separate the foreground and background for cartoonization. GANILLA [17] used unpaired cartoon images as a dataset and obtained high-quality cartoon illustration images. Chen et al. [18] converted real images into anime images, but some images still have problems such as artifacts and chromatic aberration. LI et al. [19] proposed a realistic style transfer method, but the transfer process is computationally intensive and memory intensive.

III. LWCOMICGAN

In the proposed LWComicGAN algorithm, it uses VGG-19 for feature extraction. We propose a lightweight generator network which greatly reduces the number of training parameters. A new optional normalization function residual block is designed and spectral normalization is used in the discriminator. In order to ensure the stability of training and the correct direction of gradient update, a gradient penalty system is applied to increase the generalization ability of the model. The loss function of the discriminator is used as the target loss function of the WGAN−GP loss function, and the content loss, gray loss and color loss are also added.

A. Data Preprocessing

VGG uses a Gram matrix on each layer of the network to represent the correlation between different filter groups, and each element in the matrix can describe the correlation between channels. The i-th convolutional layer in the network will have a filter bank to activate the feature information obtained by the upper layer. Each filter bank will have N_i feature map with a size of M_i according to the setting of each convolutional layer. The size of the Gram matrix G_i can be expressed as $N_i \times N_i$, h and w are the height and width of the feature map, and the inner product of the sum of the feature maps calculated according to the vectorized features formula is given as Eq. (1).

$$G_{jk}^{i} = \sum_{hw} x_{hwj}^{i} \cdot x_{hwk}^{i} \tag{1}$$

Where G_{jk}^{i} is the activation of the generated image at the (j, k)th position in the filter.

B. Loss Function

In our work, we use the loss function of the WAGN-GP network as the loss function of the LWComicGAN network, and add a gradient penalty strategy. It can well restrain the occurrence of gradient explosion and gradient disappearance. The adversarial loss formula is given as Eq. (2).

$$\begin{aligned} L_{gan} = {} & E_{\tilde{x}\sim Q_g}\left[F_\omega(\tilde{x})\right] - E_{x\sim Q_r}\left[F_\omega(x)\right] + \\ & \lambda\left[\left(E_{x\sim Q_{\hat{x}}}\left[\left\|\nabla_{\hat{x}} F_\omega(\hat{x})\right\|\right]-1\right)^2\right] \end{aligned} \tag{2}$$

Where $\tilde{x}$ is a random real sample, Q_g is a set of real samples; x is a random generated sample, and Q_r is a set of generated samples. $F_\omega(\tilde{x})$ means that under the adjustment of the parameter ω belonging to a certain range, the partial derivative of the non-linear activation layer discriminator network F_ω to the input sample $\tilde{x}$ or x is not greater than the Lipschitz constant, and the parameter ω is as long as it belongs to a certain range, no matter what the constant is, the range of F_ω doesn't exceed it definitely, and the update direction of the function to the gradient doesn't change.

The content loss is calculated by the conv4-4 layers of VGG, which can make the generated image retain the content of the input image, and the content loss formula is given as Eq. (3).

$$L_{con}(W,U) = E_q\left[\left\|VGG(W_q) - VGG(q)\right\|_1\right] \tag{3}$$

Where q represents the image in the training set of real photos, W_q is the fake image generated by the generator, $VGG(W_q)$ and $VGG(q)$ are the feature maps output after being processed by the VGG network, and the whole calculation is completed and processed by L1 regularization.

The grayscale loss preserves the line texture of the styled image, presenting a clear anime style in the resulting image. The grayscale loss formula is given as Eq. (4).

$$\begin{aligned} L_{gra}(W,U) = {} & E_q, E_x\left[\left\|Gram\left(VGG\left(W_q\right)\right)\right.\right. \\ & \left.\left. - Gram\left(VGG(q)\right)\right\|_1\right] \end{aligned} \tag{4}$$

Where x represents the images in the anime training dataset, and the Gram matrix is used to represent the feature maps produced by VGG.

The color loss makes the generated image retain the color of the real-world image, and the color reconstruction loss formula is given as Eq. (5).

$$\begin{aligned} L_{col}(W,U) = {} & E_p\left[\left\|Y(W(q)) - Y(q)\right\|_1\right. \\ & + \left\|M(W(q)) - M(q)\right\|_H \\ & \left. + \left\|N(W(q)) - N(q)\right\|_H\right] \end{aligned} \tag{5}$$

Y, M and N represent the three channels of the YUV format.

In this paper, we assign them different weights as the total discriminator loss during the computation. The total loss formula is given as Eq. (6).

$$L_{loss}(W,U)=\alpha L_{gan}(W,U)+\beta L_{con}(W,U)+\gamma L_{gra}(W,U)+\omega L_{col}(W,U) \quad (6)$$

Where α , β , γ , ω are the weight coefficients of gan loss, content loss, gray loss, and color loss.

C. Network Structure

This paper proposes a self-adjusting instance-level normalization function and designs a new residual block structure. When layer normalization is used alone, due to Layer Normalization (LN) has some defects, it combines each channel of the image for normalization, and its pixel set integrates all the images on the image. Therefore, performing LN in the case where the color of a certain picture is relatively large will cause a single color in the whole image, resulting in serious artifacts. Since Instance Normalization (IN) only focuses on pixels on a single feature map, it has nothing to do with the number of channels. It can effectively reduce the generation of artifacts. However, LN preserves the content of the source image better, and is of great help to the average stylization style of each image, we use an optional normalization function to complete the image generation.

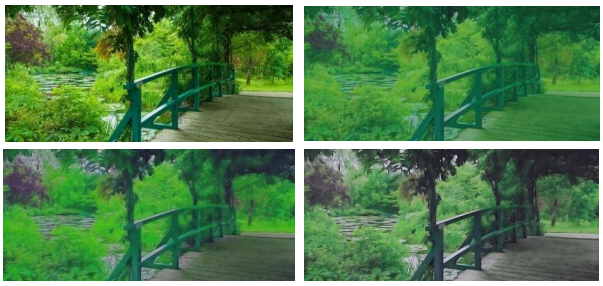

Fig. 1. Images produced using layer normalization for 20 epochs (top right), 20 epochs (bottom left) and 60 epochs (bottom right) using the model with the conditioning function, preserving the artifact-removed image effect.

We use the results of data normalization to judge the correlation with the integrated value of the input vector dimension, so as to choose an appropriate normalization function. The total loss formula is given as Eq. (7).

$$\alpha_I=\frac{x-\mu_I}{\sigma^2_I},\alpha_L=\frac{x-\mu_L}{\sigma^2_L} \quad (7)$$

Where x represents the input tensor, μ represents the mean, and σ^2 represents the variance. The input tensor represents the feature data of the image, and subtracting the mean value from the data can highlight the differences between different pixels of the image. These differences are passed to the neural network to help the network judge the features. Dividing by the variance is to make the network pay attention to the pixel value changes of the entire image, eliminate the local change gap, and make the pixel changes more balanced. Then compare the values of α_I and α_L and use the normalization function to which the larger value belongs as the normalization function required for this layer.

In this paper, we design a lightweight framework. It consists of convolutional layers, four residual blocks and upsampling modules. The first convolutional layer that processes the input image has a 7×7 kernel. The remaining convolutional layers have a 3×3 kernel, and each convolutional block uses a Leaky Relu activation function. The last layer has only one convolutional layer followed by a tanh activation function that finally outputs the image. The structure of the generator is shown in Figure 2.

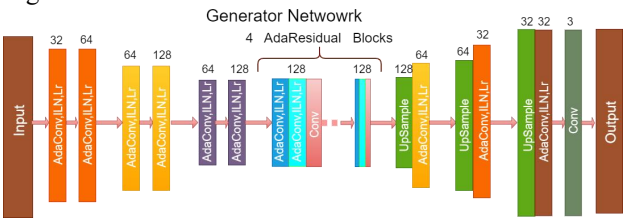

Fig. 2. Generator network diagram.

A residual block with an optional instance layer normalization function is designed to replace the ordinary residual block in ResNET. We replace the normalization layer with LIN layer. The residual block structure is shown in Figure 3.

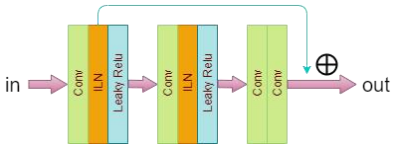

Fig. 3. The residual block structure.

In order to make the discriminator avoid the gradient explosion problem and avoid the data fed back to the generator from being too extreme. This paper uses Spectral Normalization in the discriminator. The discriminator structure is shown in Figure 4.

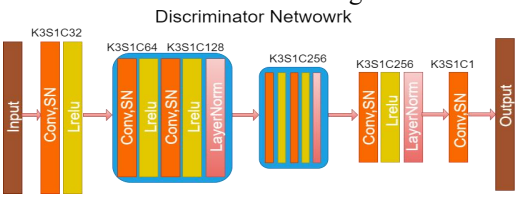

Fig. 4. Discriminator network structure diagram.

IV. Experimental Results and Analysis

A. Experiment Platform And Dataset

The experimental configuration used in this algorithm is Intel(R) Xeon(R) Gold 8-core processor. The GPU uses GeForce RTX 2080Ti with 11GB of video memory, and the experiment uses tensorflow-gpu with deep learning platform version 1.14.0. The training dataset is divided into two parts. The first part is a real photo dataset with a total of 6656 real landscape photos. The

second part is a style dataset made from the anime movie "The Wind Rises", with a total of 1792 images. The specific parameter data are shown in Table 1.

TABLE I. THE PARAMETER COEFFICIENTS IN (6)

Parameter	α	β	γ	ω
Value	350	2.5	12.5	15

In order to realize the photos with the style of Chinese ink animation. We select the famous ink animation film "Landscape Emotion" as the data source of the experiment. The dataset is made by automatic frame interception written in python, with a total of 718 pictures, and the size of the pictures in the dataset is 256×256 pixels. The image production of the dataset is shown in Figure 5.

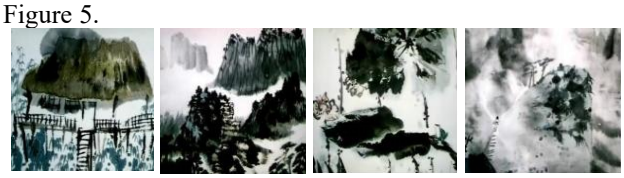

Fig. 5. Sample dataset image.

B. Experimental Results

In this experiment, animation style transfer is performed on the "The Wind Rises" dataset, which can turn real photos into images of the specified dataset animation style. In order to make a comparison, the dataset is applied to the VGG, CycleGAN and WCT [21] algorithms in the experiment. The algorithm comparison images are shown in Figure 6.

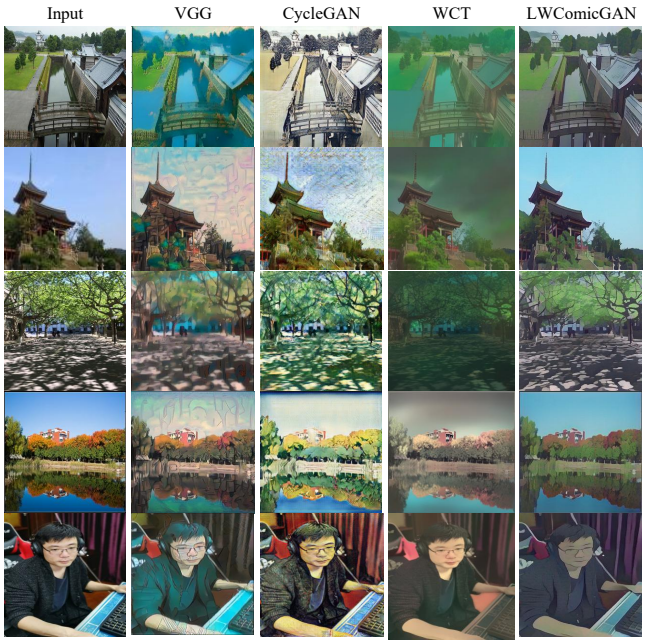

Fig. 6. Algorithm images comparison diagram.

Based on the datasets of "Landscape Emotion" and "The Wind Rises". We make a realistic photo transfer in ink painting style. The ink images are shown in Figure 7.

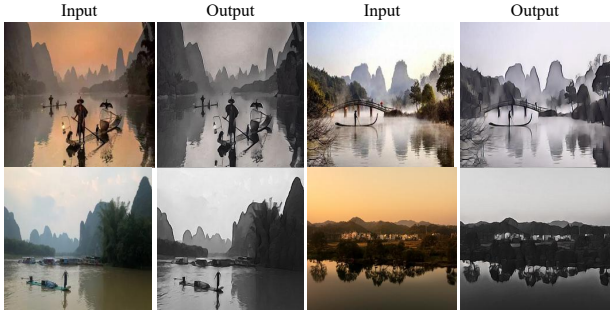

Fig. 7. Ink animation style generation diagram.

C. Experiment Analysis

There are usually two indicators to measure the quality of the image generated by the GAN network, IS and FID. IS only considers the quality of the samples generated by GAN, and FID not only considers the quality of the generated images, but also integrates the Frechet Inception distance with the original image feature vector. In order to prove the generation effect of this experiment, two methods are selected as verification criteria. IS takes into account the clarity and diversity of the image itself. The larger the value, the better the image quality. The FID value represents the variable distribution distance between the generated image and the original image. The smaller the FID value, the better the image generation quality.

TABLE II. FID AND IS SCORE COMPARISON

Method	VGG	CycleGAN	WCT	LWComicGAN
FID	338.22	202.15	193.86	173.80
IS	6.32	8.48	6.90	10.24

CartoonGAN and AnimeGAN [22] have great achievements in style transfer, but the amount of parameters they generate is indeed huge. The comparison of the total number of parameters in the training process with them are shown in Table 3.

TABLE III. TOTAL PARAMETER COMPARISON

Method	CartoonGAN	AnimeGAN	LWComicGAN
Parameter Size	13253452	3956096	2796416

V. CONCLUSIONS

In this paper, we improves the image generation algorithm based on traditional GAN network and proposes a lightweight animation transfer algorithm LWComicGAN, which reduces the number of parameters. Traditional methods are difficult to maintain the integrity

of image content. The training is unstable, and the animation effect is not obvious. In this paper, a variety of loss functions are added to realize the animation transfer of complete photos. The " Landscape Emotion" dataset with a total of more than 700 pictures has been produced, and the effectiveness of the algorithm in this paper has been proved on the new dataset.

In future work, more effective loss functions should be comprehensively applied to further improve the network structure.

ACKNOWLEDGMENT

This work was supported by the National Natural Foundation of China (E040101, 50811120111, 51574221, 41874044); Shandong University of Science and Technology (Taian) Scientific Research Innovation Team Project (2013KYTD04); Shandong University of Science and Technology Scientific Research Platform Project (2014KYPT30); Qingdao Philosophy and Social Science Planning Project (QDSKL2101139).

REFERENCES

[1] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Gupta, Agrim, et al. "Characterizing and improving stability in neural style transfer." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[3] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." European conference on computer vision. Springer, Cham, 2016.

[4] Cui, J., et al. "PortraitNET: Photo-realistic portrait cartoon style transfer with self-supervised semantic supervision." Neurocomputing 465 (2021): 114-127.

[5] Castillo, Carlos, et al. "Son of zorn's lemma: Targeted style transfer using instance-aware semantic segmentation." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

[6] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[7] Chen, Yang, Yu-Kun Lai, and Yong-Jin Liu. "Cartoongan: Generative adversarial networks for photo cartoonization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[8] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[9] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[10] Zhao H H, Rosin P L, Lai Y K, et al. Image neural style transfer with global and local optimization fusion[J]. IEEE Access, 2019, 7: 85573-85580.

[11] J. Lian and J. Cui, "Anime Style Transfer With Spatially-Adaptive Normalization," 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428305.

[12] Gong J, Hold-Geoffroy Y, Lu J. Autotoon: Automatic geometric warping for face cartoon generation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 360-369.

[13] T. Wei and L. Zhu, "Comic style transfer based on generative confrontation network," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 1011-1014, doi: 10.1109/ICSP51882.2021.9408938.

[14] Wang X, Yu J. Learning to cartoonize using white-box cartoon representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8090-8099.

[15] Y. Shu et al., "GAN-based Multi-Style Photo Cartoonization," in IEEE Transactions on Visualization and Computer Graphics, doi: 10.1109/TVCG.2021.3067201.

[16] Way D L, Chang W C, Shih Z C. Deep learning for anime style transfer[C]//Proceedings of the 2019 3rd International Conference on Advances in Image Processing. 2019: 139-143.

[17] Hicsonmez S, Samet N, Akbas E, et al. GANILLA: Generative adversarial networks for image to illustration translation[J]. Image and Vision Computing, 2020, 95: 103886.

[18] Chen Y, Lai Y K, Liu Y J. Cartoongan: Generative adversarial networks for photo cartoonization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9465-9474.

[19] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 172-189.

[20] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[21] Li Y,Liu M Y, Li X, et al. A closed-form solution to photorealistic image stylization[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 453-468.

[22] C Gulrajani hen J, Liu G, Chen X. AnimeGAN: A novel lightweight gan for photo animation[C]//International Symposium on Intelligence Computation and Applications. Springer, Singapore, 2019: 242-256.