2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)

2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA) | 978-1-6654-5911-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVIDLICCEA56201.2022.9824311

Trans-CycleGAN: Image-to-Image Style Transfer with Transformer-based Unsupervised GAN

Shiwen Li
Xidian University, Xi'an, China, 710000
* Corresponding author: 21171213866@stu.xidian.edu.cn

***Abstract*—The field of computer image generation is developing rapidly, and more and more personalized image-to-image style transfer software is produced. Image translation can convert two different styles of data to generate realistic pictures, which can not only meet the individual needs of users, but also meet the problem of insufficient data for a certain style of pictures. Transformers not only have always occupied an important position in the NLP field. In recent years, due to its model interpretability and strong multimodal fusion ability, it has also performed well in the field of computer vision. This paper studies the application of Transformers in the field of image-to-image style transfer. Replace the traditional CNN structure with the improved Transformer of the discriminator and generator model of CycleGAN, and a comparative experiment is carried out with the traditional CycleGAN. The test dataset uses the public datasets Maps and CelebA, and the results are comparable to those of the traditional CycleGAN. This paper shows that Transformer can perform the task of image-to-image style transfer on unsupervised GAN, which expands the application of Transformer in the CV filed, and can be used as a general architecture applied to more vision tasks in the future.**

Keywords- CycleGAN; Transformer; CNN; Image-to-image Style Transfer

I. INTRODUCTION

CycleGAN(Cycle-Consistent Generative Adversarial Networks) is an unsupervised, conditional GAN [1]. Image-to-image style transfer can be performed with unpaired data, which is characterized by capturing the style features of the image on a dataset, and then converting these features to another type of style dataset to complete the image-to-image style transfer. Currently, CNN occupies an important position in the field of CV, and recently, Transformer, which plays an important role in the field of NLP, has also begun to perform well in classification, image and video recognition tasks. In particular, the Vision Transformer has been proven in the field of CV to achieve good classification accuracy on ImageNet with less computing power Flops [2]. This paper investigates whether CycleGAN can use Transformer to improve the generator and discriminator, using pure Transformer structure model, to achieve the task of image-to-image style transfer of unsupervised GAN. Through the experiments in this paper, two VIT-based generators and discriminators are designed and trained and tested on the small size dataset Maps and the medium size dataset CelebA. The experiments show that the Transformer-based CycleGAN model is comparable to the traditional CNN-based The CycleGAN model is equivalent, but more stable and general. Transformer can be used as a general architecture on unsupervised GAN.

II. RELATED WORK

A. Related Research Of CycleGAN

GAN is composed of a generator and a discriminator. It adopts the idea of game theory, and the generator simulates the data distribution of real images to generate images. The discriminator is used to make judgments between the generated image and the real image. The continuous optimization between the two networks finally generates images that can be faked. CycleGAN is based on the idea of double learning, learning the unpaired dataset between the source domain and the target domain, and transforms the image style. Firstly transform from the source domain to the target domain, and then from the target domain to the source domain, and ensure the consistency of the final generated image with the source domain image. Using cycle consistency to solve the problem of image conversion, the generator will also continue to optimize iterations due to the results of the discriminator. Image-to-image style transfer can be traced back to the Image Analogy model proposed by Herzmann et al. [3], which employs a non-parametric texture model on a single input-to output training image. Later, more methods use sample datasets to train convolutional neural networks. This paper uses an unsupervised generative adversarial network to learn a mapping from input images to output images. Similar ideas have been applied to a variety of different tasks with remarkable results. For example, generate images from contours [4], image attributes and semantic layouts [5].

B. Transformer Related Works And Progress

In recent years, Transformer models have achieved remarkable success in CV, such as ViT(Vision Transformer) and DeiT [6]. ViT uses images as a sequence of tokens, and uses a pure Transformer model structure to replace the convolution and pooling operations in the CNN. It is suitable for image classification tasks and achieves good results with lower computational costs on ImageNet. Due to the excellent prospects of Transformer in modeling non-local context dependencies, it shows its scalability and high efficiency, so after the ViT is generated, it has been used in target detection, video recognition, multi-task pre-training and other tasks. And there is already work on applying Transformer to GAN, mainly for image generation task. Inspired by CNN, Parmar et al. [7] proposed Image Transformer, which was the first to apply the Transformer to image transformation and image generation tasks. It employs an image generation formulation

978-1-6654-5911-2/22/$31.00 ©2022 IEEE

similar to sequence modeling within the Transformer framework. The model consists of two parts, an encoder for extracting image representations and a decoder for generating pixels. However, this model only focuses on the local attention range, and the image generation depends on the surrounding values of each pixel, which has high storage and computational costs. Esser et al. [8] proposed VQGAN, which combines the effective inductive bias of CNN and the expressive power of Transformer, the first Transformer architecture to generate megapixel images guided by semantics. The authors use a CNN architecture to model image components, and a Transformer architecture to synthesize the components, representing the image as composed of perceptually rich image components, resulting in a high resolution image. VQGAN does not need to relearn the known and regular knowledge of the local structure of the image, and maintains the flexibility of Transformer while effectively encoding the inductive bias. Last year, Yifan Jang et al. proposed TransGAN [9], using a pure Transformer structure for the task of image generation, which was the first work using Transformer on GAN. Structurally, TransGAN consists of two parts: the generator can gradually increase the feature resolution while reducing the embedding dimension at each stage; the other is a patch-level discriminator, which takes image patches as input instead of pixels, and classify between real and generated images. Meanwhile, a multi-task co-training strategy along with a locally initialized self-attention mechanism is used to enhance the neighborhood smoothness of natural images. Experimental results show that TransGAN achieves good results on the CIFAR10 dataset, but slightly lower than StyleGAN v2 [10] on the larger scale and higher resolution STL-10 benchmark. It is therefore concluded that pure Transformers are sufficiently capable for difficult image generation tasks. There is also ViTGAN [11] proposed by Kwonjoon Lee et al., which proves that ViT has comparable results to CNN structures on the image-to-image generation task of GAN.

III. Method

A. The Structure of CycleGAN

The model of CycleGAN completes the mapping from the source domain to the target domain, and then converts back from the target domain, which removes the limitation of requiring paired dataset. At the same time, the constraints of cycle consistency are added to solve the problems that may arise from the above-mentioned traditional GANs. CycleGAN is similar in structure to a ring network structure, as shown in Figure 1.

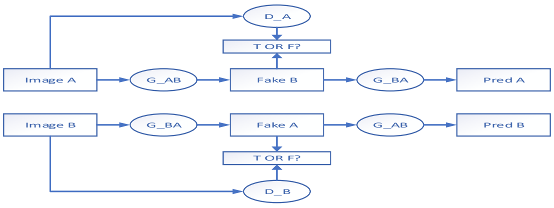

Figure 1. CycleGAN Structure

B. The improved CycleGAN

The model consists of two generators $G: X \rightarrow Y$ and $F: Y \rightarrow X$, and two discriminators D_X and D_Y. $D_X$is used to distinguish $F(y)$ generated by dataset $\{x\}$ and F, and D_Y is used to distinguish $G(x)$ generated by dataset $\{y\}$ and G. CycleGAN's full loss consists of three parts. The Adversarial loss, is used to make the image generated by the generator closer in distribution to the target domain Y. The cycle consistency loss is used to prevent the generative network from training too well so that the output image matches the target domain, but it may not match the input image. The identity loss helps to preserve the color of the input image, and it is expected that the generated image will be used as input again, and the same result can be obtained. Compared with other CycleGAN, the most significant difference between our method is its unique generator and discriminator. This paper proposes a improved structure using Transformer as part of image understanding, including two Trans-based discriminators and two Trans -based generators.

For the mapping function G and the discriminator D_Y, the objective function is:

$$L_{GAN}(G, D_Y, X, Y) = E_{y\sim Pdata(y)}[\log D_Y(y)] + E_{x\sim Pdata(x)}[\log(1 - D_Y(G(x)))] \tag{1}$$

The purpose of G is to minimize the objective function, while the purpose of D_Y is to maximize the objective function:

$$\min G \max D_Y L_{GAN}(G, D_Y, X, Y) \tag{2}$$

The same adversarial loss function is used for the mapping function F and the discriminator D_X:

$$\min F \max D_X L_{GAN}(F, D_X, Y, X) \tag{3}$$

The cycle-consistency loss of CycleGAN is set as:

$$L_{cyc}(G,F) = E_{x\sim Pdata(x)}[\|F(G(x)) - x\|_1] + E_{y\sim Pdata(y)}[\|G(F(y)) - y\|_1] \tag{4}$$

The Identity loss is set as:

$$L_{idt}(G,F) = E_{y\sim Pdata(y)}[\|G(y) - y\|_1] + E_{x\sim Pdata(x)}[\|F(x) - x\|_1] \tag{5}$$

So the full loss function is obtained as:

$$\begin{aligned} L(G,F,D_X,D_Y) &= L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,Y,X) \\ &+ \lambda L_{cyc}(G,F) + \lambda \times \lambda_{idt} L_{idt}(G,F) \end{aligned} \tag{6}$$

λ and λ_{idt} are the weights of the loss item in the full loss function.

C. Improved Trans-based Generator

The generator in this paper consists of two parts, a transformer block and an output mapping layer. VIT needs to divide the image into patches, it is expected to turn the discrete input into a continuous signal representation, use a continuous function to represent the real state of the image, and use a neural network to approximate this continuous function. We need to find a function that satisfies certain constraints. This paper uses Implicit Neural Representation to fit the data, learn a continuous mapping from patch embedding to pixel values, and penalize points that don't satisfy constraints. Choosing a periodic function such as sin as the activation function makes the representation form smoother, fits the gradient and image well, and makes the image clearer. Several experiments have proved that the representation ability of the SIREN network is stronger than that of the fully connected network with ReLU as the activation function [12]. Part of the code for the generator structure is as follows:

```
x = self.mlp(z).view(-1, self.initialize_size * 8, self.dim)
x, h=self.Transformer_Encoder(self.position_emb, x)
x = self.SLN(h, x)
x = self.INR(x)
result=x.view(x.shape[0],3,self.initialize_size*8,self.initiali
ze_size * 8)
```

The result of using Self-modulated LayerNorm after positional encoding is used as the input of the transformer block. Through Implicit Neural Representation to learn the continuous mapping from patch embedding to pixel.

$$\mathrm{SLN}(h_t, w) = \mathrm{SLN}(h_t, \mathrm{MLP}(z)) = \gamma_t(w) \odot \frac{h_t - \mu}{\sigma} + \beta_t(w) \qquad (7)$$

Where the variance and mean are controlled by μ and σ, and the adaptive normalization parameter W of the latent vector from Z is controlled by γ_1 and β_1.

D. Improved Trans-based Discriminator

The traditional CycleGAN uses PatchGAN as the discriminator of the network, instead of taking the whole image as the input of the discriminator, but dividing the image into patches as input. The method in this paper learns this idea. After dividing the image into patches, each patch is compressed into tokens through the network, and then combined with the learnable positional encoding to form a sequence, and then cls-token is added to the first position of the sequence. The output of the cls-token is classified by the Transformer encoders. Continuity plays a key role in GAN's discriminator. After satisfying the Lipschitz continuity in WGAN, the entire network can be trained very stably. However, a recent work shows that Lipschitz continuity is violated in the Self-attention mechanism. MLP networks are Lipschitz continuous, but Dot-product multi-head self-attention is not Lipschitz continuous. The improved L2 multi-head self-attention is Lipschitz continuous. This paper uses a self-attention operation based on L2 norm proposed in [13] to improve the stability of the discriminator in GAN. And use spectral normalization to further strengthen the continuity. Part of the code for the discriminator structure is as follows:

```
img_patch = torch.cat((cls_token, img_patch), dim=1)
img_patch = img_patch + self.position_emb[: tokens + 1, :]
img_patch = self.emb_dropout(img_patch)
result = self.Transformer_Encoder(img_patch)
log = self.mlp_head(result[:, 0, :])
log = nn.Sigmoid()(log)
```

The attention used is as follows, that is, using a Query, calculate the similarity between it and each Key and use it as a weight, and then perform a weighted sum of all Values:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \qquad (8)$$

IV. Experiment

A. Dataset

Maps is a small public dataset. This paper uses 2194 aerial images and 2194 maps as a training set. The aerial images and maps are unpaired images. The CelebA dataset, which is a public dataset, contains a large number of images of male and female faces. The files are saved to the male and female folders according to gender, including 84,434 male pictures and 118,165 female pictures. This paper completes a gender style transfer experiment between men and women.

B. Experiment Process

The experiment sets the input image size to 64×64 resolution and 128×128 resolution. In the experiment, the epoch is set to 10000, the learning rate is set to 0.0002 for the first 25000 epochs, the λ_A and λ_B are set to 10, and the λ_{idt} is set to 0.5. A series of data argumentation such as resize and random horizontal flip are performed on the input image. During the training process, the Adam optimizer is used to optimize the gradient descent.

The input image is divided into patches through patch embedding, and the feature dimension is reduced to 384 through linear, and 4 attention heads are designed.

For adversarial loss, using the least squares function instead of the negative log-likelihood function can make the training results more stable [14], and can also make the quality of the generated images higher. In order to reduce the oscillation of the model, before the images are input to the discriminator, the image buffer is used to store 30 images, which are randomly updated as the epoch increases, and the newly generated images are input to the discriminator to calculate the loss.

C. Main Results

This paper uses PSNR and SSIM as evaluation metrics. PSNR calculates the difference between the input image and the output image, estimating the quality of the picture. SSIM is used to measure how similar the input image and output image are. After 100 real Maps pictures and face pictures are tested, the average value is calculated.

TABLE I. RESULTS ON MAPS 64*64

	PSNR	SSIM
The CNN-based CycleGAN	26.748	0.614
The Trans-CycleGAN	26.716	0.596

TABLE II. RESULTS ON CELEBA 128*128

	PSNR	SSIM
The CNN-based CycleGAN	24.425	0.534
The Trans-CycleGAN	24.459	0.609

It can be seen that on the small dataset Maps, the effect is slightly worse than the traditional CycleGAN, however, using the model in this paper improves the PSNR by 0.034 and the SSIM by 0.075 on dataset CelebA.

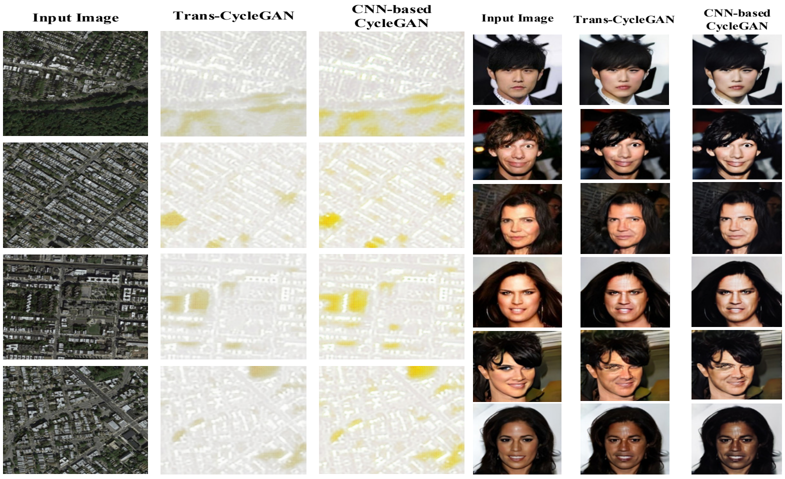

Figure 2. Results produced by Trans-CycleGAN and traditional CycleGAN

In addition to the above indicators, the effect of the image in the human eye is also important. The results are shown in the figure. It can be seen that the Transformer-based model is equivalent to the traditional CycleGAN. The Transformer-based CycleGAN model in this paper completely discards the convolution operation and completely uses self-attention to learn the inductive bias, which can achieve performance comparable to the CycleGAN using convolution operations, but is more stable and versatile. It is proved that the full Transformer structure can be used to replace the generator and discriminator of the traditional CNN structure in CycleGAN.

V. Conclusion

This paper proposes an improved CycleGAN model based on Transformer, The model achieves comparable performance to CNN-based CycleGAN on Maps and CelebA. It shows that the Transformer can perform the task of image style transfer on the unsupervised GAN, which expands the application of the Transformer in the CV , so that it can be used as a general architecture for various tasks of CV. However, compared with CNN, Transformer has a greater demand for the amount of input data, resulting in a very large amount of training calculations. For large data, Transformer can have a good effect on CycleGAN, but for small and medium data, the effect is not much different from CNN. It is hoped that more Transformers can be applied to different tasks in the future, and it is expected to become a general architecture in the CV field.

ACKNOWLEDGENMENT

Thanks to Prof. Lio', I was exposed to GAN and Transformer, which resulted in the topic selection of the thesis. We would like to thank Mr. Wang for her careful guidance and guidance from the conception to the final draft of this paper. I would like to thank my family for their support in my academic research.

References

[1] Jun-Yang Zhu, T. Park, P. Isola, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks [J]. IEEE, 2017.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [J]. 2020.

[3] A. Hertzmann, C. E. Jacobs, N. Oliver, et al. Image analogies [C]. International Conference on Computer Graphics and Interactive Techniques, 2001: 327-340.

[4] [4] P. Sangkloy, J. Lu, C. Fang, et al. Scribbler: Controlling deep image synthesis with sketch and color [C]. Computer Vision and Pattern Recognition, 2017: 6836-6845.

[5] L. Karacan, Z. Akata, A. Erdem, et al. Learning to generate images of outdoor scenes from attributes and semantic layouts [J]. arXiv: Computer Vision and Pattern Recognition, 2016.

[6] H. Touvron, M. Cord, M. Douze, et al. Training data-efficient image transformers & distillation through attention [J]. 2020.

[7] N. Parmar, A. Vaswani, J. Uszkoreit, et al. Image Transformer [J]. 2018.

[8] P. Esser, R. Rombach, B. Ommer. Taming Transformers for High-Resolution Image Synthesis [J]. 020.

[9] Y. Jiang, S. Chang, Z. Wang. TransGAN: Two Transformers Can Make One Strong GAN [J]. 2021.

[10] T. Karras, S. Laine, M. Aittala, et al. Analyzing and Improving the Image Quality of StyleGAN [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.

[11] K. Lee, H. Chang, L. Jiang, et al. ViTGAN: Training GANs with Vision Transformers [J]. 2021.

[12] V. Sitzmann, J. Martel, A. W. Bergman, et al. Implicit Neural Representations with Periodic Activation Functions [J]. 2020.

[13] H. Kim, G. Papamakarios, A. Mnih. The Lipschitz Constant of Self-Attention [J]. 2020.

[14] X. Mao, Q. Li, H. Xie, et al. Least Squares Generative Adversarial Networks [J]. IEEE, 2017.