CONVERTING REAL HUMAN AVATAR TO CARTOON AVATAR UTILIZING CYCLEGAN

Wenxin Tian

Graduate School of Information Sciences and Arts,
Toyo University, Kawagoe, Saitama, Japan
Dept. of Information Sciences and Arts, Toyo University, Kawagoe, Saitama, Japan

ABSTRACT

Cartoons are an important art style, which not only has a unique drawing effect but also reflects the character itself, which is gradually loved by people. With the development of image processing technology, people's research on image research is no longer limited to image recognition, target detection, and tracking, but also images In this paper, we use deep learning based image processing to generate cartoon caricatures of human faces. Therefore, this paper investigates the use of deep learning-based methods to learn face features and convert image styles while preserving the original content features, to automatically generate natural cartoon avatars. In this paper, we study a face cartoon generation method based on content invariance. In the task of image style conversion, the content is fused with different style features based on the invariance of content information, to achieve the style conversion.

KEYWORDS

Deep learning, CNN, Style transfer, Cartoon style.

1. INTRODUCTION

Cartoon faces appear in animations, comics, and games. They are widely used as profile pictures in social media platforms, such as Facebook and Instagram. Drawing a cartoon face is labor intensive. Not only it requires professional skills, but also it is difficult to resemble unique appearance of each person's face. Through style transfer, which can express the picture effect more perfect and be able to achieve the desired effect. It can be done without complex PS retouching and does not require particularly good drawing skills to complete the corresponding task. In the film production or webcast, it can make the image performance more involved, more vivid image special effects to do more abstract perfection. Image stylization originated from the research of Gatys and others, who found that although today's style migration has achieved good results, there are still some areas for improvement. The first thing to solve is the time consuming problem, even if you choose the optimal solution, it takes a long time to train a model, obviously there is still a lot of room for improvement, and there is still a lot of room for optimizing the time problem for selfie images.

In recent years some social networking services have been popular such as TikTok. Photo-to-cartoon style transfer for face can be useful for the services especially when the users do not want to show their own faces. And due to COVID-19, many schools have adopted online classes to prevent the expansion of the infection. Teachers want to know how well their students understand ,what they learned and how well the students focus on. What the teachers said from nonverbal information such as facial expression, facial pose, eye-gaze, etc. On the other hand many students do not want to show their faces. In this case photo-to-cartoon style transfer can be

DOI: 10.5121/ijaia.2022.13601

useful, because it can keep facial expression, facial pose, and eye-gaze, while it converts real photo style into cartoon style. The purpose of this paper is to construct a model which can convert real face images into cartoon face images using deep neural networks.

2. BACKGROUND AND SIGNIFICANCE

Face cartoon caricature generation is a highly interesting research in the field of image processing. Due to its unique expression and diverse styles, cartoon cartoons are gradually becoming popular among people. Under the influence of cartoon cartoons, companies are using cartoon characters or animals as their spokespersons, and people are designing personalized cartoon avatars for their social accounts. People design personalized cartoon avatars for their social accounts. Traditionally, cartoon avatars are usually obtained by hand-drawing. This method takes a lot of time to compose, trace, and color, and is quite expensive. Although this method can be used to obtain the most detailed and natural cartoon avatar, it is not suitable for today's era of rapid technological development, and therefore As a result, a large number of personalized software has been created. Thus, cartoon caricatures are not only widely used in people's life The cartoon caricature is not only widely used in people's life and entertainment, but also brings great fun to people. Therefore, the study of face cartoon cartoon generation technology The study of face cartoon generation technology has important research significance and application value.

The images obtained by the cartoon avatar customization software are mainly composed of multiple face parts, so the similarity with the original image is low, and the real sense of personalization should be personalized for the face image design. With the development of image processing technology, the generation of face cartoon images can be realized with the help of style migration technology. Face style migration mainly uses image processing technology to fuse the content of the original image with the style of the style image to achieve the conversion of style. The automatic generation of cartoon images can largely reduce the workload of painters and get stylized images quickly. Nowadays, thanks to the continuous research on this technology, some similar applications have appeared on the market, which are mainly based on image processing methods and are able to convert face images into corresponding cartoon images and, depending on the drawing style, into different styles of cartoon faces. It can be seen that the study of face cartoon cartoon generation methods based on image processing has become a trend. In the traditional face cartoon method, part of it is mainly based on the contour information of facial features to generate simple line The other part is based on the machine learning method, which can build a face cartoon by simple learning of the sample. The other part is based on the machine learning method, which establishes the matching relationship between the face image block and the cartoon image block by simple learning of the sample, and finds the cartoon image block that best matches the original image block. The other part is based on machine learning, which establishes the matching relationship between the face image block and the cartoon image block by simply learning the sample, finds the cartoon image block that best matches the original image block, and then synthesizes the cartoon head. Although this method can achieve style The problem of unnatural expression and single effect exists. It also has the problems of unnatural expression and single effect, and the processing steps are complicated and less efficient. Deep learning is a new branch of machine learning. Because of its powerful learning ability, algorithms based on deep learning have good improvements in performance and have played a key role in the research related to image processing field, such as image classification, image segmentation, target detection and tracking, etc. In addition, convolutional neural networks can realize image to image conversion by automatic coding and decoding of images, thus enabling end-to-end conversion. image and image interconversion by automatic image coding and decoding, thus enabling end-to-end image style conversion, which makes automatic cartoon generation possible. This makes it possible to automatically generate cartoons. However, due to

the special nature of face images, there are large differences and uniqueness between individuals. The style conversion of face images needs to be further researched, especially for the five facial features.

Therefore, further research is needed for the style conversion of face images. All in all, it is a challenging problem to achieve automatic style conversion of face images in a fast and high quality way. In summary, the research on the automatic generation method of face cartoon and exaggerated cartoon has important scientific significance and application value.

3. CURRENT STATUS OF RESEARCH ON THE TOPIC

For face cartoon, the most important thing is to ensure the conversion of the image style, face cartoon is basically in accordance with the shape of the real human face features for painting, through different color expression and diverse styles to present different cartoon effect. Different styles of images have different artistic expressions, and the so-called style conversion is to convert the image from one style to another. In the field of image processing, the study of face images has been a hot topic, and the study of methods to generate different styles of cartoon avatars from real human face images has also received keen attention. Through years of research by scholars, image stylization has made a major breakthrough and can be applied to the task of face image stylization. The existing methods of face style conversion mainly generate cartoons based on the information of human face shape and five senses outline, these methods need to collect art style images as reference samples, and then convert the original images into images with a specific style, and these methods can be divided into two kinds, one is to directly chunk the images or separate the important five senses parts, and somehow match with the image blocks in the database to get the most similar images. The other method is to first learn a high-dimensional implicit space through a deep learning model, map the original image or style image into this space, and then restore the style through the corresponding decoding network to realize the conversion from real image to style image, which is called the image conversion task, and apply this This is called image conversion task, and applying this method to face images, we can also realize face stylized conversion. Therefore, the current status of face stylization research is described below: face stylization based on deep learning.

Face stylization based on automatic generation is also generally based on a learning approach, where a conversion model is obtained by learning from sample instances, and a corresponding cartoon image is automatically generated by inputting real face images into the conversion model. Deep learning, as a new field in computer vision, has shown good performance on many tasks, such as image classification, target detection and semantic segmentation. With the gradual development of deep learning, direct image-to-image conversion becomes possible, and this task is called image translation. The style migration method based on convolutional neural network was first proposed by A. Gatys. This method mainly extracts the style of the style image by pre-trained deep neural network, then extracts the content of the original image, and finally superimposes the extracted style with the content to get an image with a specific style. However, this method can only convert the overall style of the image and change the overall texture features of the image, while the cartoon face image focuses on the style conversion of the face part. Therefore, this method does not generate a cartoon face image in the true sense.

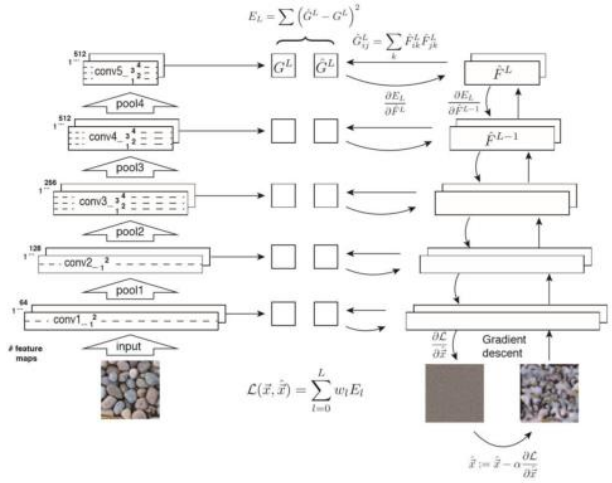

Figure 1. Image texture extraction based on CNN

In recent years, generative adversarial networks have made great progress in the field of image generation, and as the name implies, the network consists of two sub-networks, the generative network and the adversarial network. The original generative network is used to implement the noise-to-image conversion, while the discriminative network is used to determine whether the image generated by the generative network is true and belongs to the same style or the same class as the target image: true is 1 for true style images and false is 0 for images generated by the generative network. The core idea of the network is to make the distribution of generated images gradually converge to the distribution of images in the target domain by using the process of mutual game between the generative network and the discriminative network. In the training process, the discriminative network is first trained to accurately determine whether the image is real, and then the generative network is trained in the hope that the generated image can fool the discriminative network, and the two networks are mutually constrained to progress together, so that the generative network can generate a fake image. The original generative adversarial network is to establish the mapping from noise to image, while the conditional generative adversarial network can realize the mapping between different contents, including text, image, and video frames.

Image conversion algorithms can be generally classified into two categories, one is generative adversarial networks based on paired data, and the other is generative adversarial networks based on unpaired data.

The first class of methods has extremely high requirements on the database, i.e., each original image should have a target style image corresponding to it, and the generative network is constrained by comparing the differences between the generated image and the real target image during the training process. The most representative method of this type is Pix2pix, the network structure of this algorithm consists of a generative network and a discriminative network. The discriminant network is essentially a binary network for discriminating whether the input image belongs to the same style as the real image. Unlike the original generative adversarial network, this method uses the original image as the input of the discriminant network in a stitching way: the original image is stitched with the generated image as a negative sample, and stitched with the real style image as a positive sample to determine whether the generated image is a real image pair, in order to discriminate the input image in terms of style on the one hand, and to ensure that the two stitched images have the same content on the other hand . However, this

method has some limitations, i.e., in real life, it is difficult to receive a large amount of paired data, which is the biggest difficulty faced by such methods.

The second type of methods emerged to solve the limitations of the first type of methods. The main principle of this type of methods is that, assuming a common implicit space exists for two styles of images, then by mapping the image of style A to the implicit space, it can be converted into an image with style B using the decoder of the corresponding style. CycleGAN uses the idea of reconstruction to constrain the generative network, in the case that there is no real image corresponding to ground truth, an effective model can make the image change from style A to style B, then it can also change from style B back to A. The reconstruction method makes the content of the image guaranteed, and this method is based on the assumption that there are two independent implicit spaces between styles A and B. This method is based on the assumption that there are two independent implicit spaces between styles A and B, i.e., the implicit space from A to B and the implicit space from B to A are independent of each other.

Based on training image translation tasks based on unpaired data, Lee H proposed that the image-to-image mapping relationship should be multimodal by nature, that is, for one input, there can be multiple outputs. Some methods solve this problem directly by adding noise, but this method lacks the constraint between noise and target distribution and is prone to pattern collapse, and although BicycleGAN constrains the input noise, the input images are still required to be paired. Based on this, the authors propose a decomposition structure, using the structure of the generative network to divide the generative network into an encoder and a decoder, and build two encoders at the same time, one is trained to obtain the style features of the stylized image, and the other is trained to obtain the content features of the original image, and then fuse these two features into the style decoder to generate an image with a unique style. A similar approach is MUNIT, with the difference that the approach uses random noise as style features in the generation phase, but constrains the input noise that the style features of the generated image should be consistent with the input random noise.

In conclusion, deep neural networks have a powerful feature representation capability, and the conditional generation adversarial network-based The style conversion method based on conditional generative adversarial network combines this feature and can easily realize the face style conversion by encoding and decoding the image features. It is the most effective method for face stylization at present. This paper is also based on the idea of generative adversarial network to realize the conversion from human face image to cartoon image.

4. RELATED RESEARCH

4.1. Unpaired Image-to-Image Translation

When paired images such as a real face image and the corresponding cartoon face image are needed to train photo-to-cartoon style transfer model, obtaining strictly paired images is very difficult. So the model that do not rely on paired images are of great practical importance.

There is a gap between paired and unpaired picture training that cannot be eliminated. Nevertheless, in many cases, it is still feasible enough to use unpaired data exclusively. Zhu et al. [2] expand the range of possible uses of "unsupervised" configurations. To some extent, it solves the problem of deep learning: too little labeled data, difficulty in finding paired data, and using unpaired datasets for training.

They proposed a method to learn from the source data domain X to the target data domain Y in the absence of paired data. Its goal is to use an adversarial loss function to learn the mapping $G: X \rightarrow Y$, making it difficult for the discriminator to distinguish the picture $G(X)$ from the picture Y. Since such a mapping is subject to huge limitations, an opposite mapping $F: Y \rightarrow X$ is added to the mapping G to make them pairwise, and a cyclic consistency loss function is added to ensure that $F(G(X)) \approx X$.

4.2. Landmark Assisted CycleGan

Wu et al. [3] proposed a method to generate cartoon faces based on input human faces by utilizing unpaired training data.

The process is divided into three main steps. First, the generator generates a rough cartoon face based on CycleGAN; afterwards, the model generates a pre-trained regression volume to predict the facial landmark based on the image generated in the first step, which marks the key points of the face. Finally, with both local and global discriminators, the researchers refine the face features in the cartoon image and the corresponding real image. In this stage, the consistency of the landmark is emphasized, so the final generated results are realistic and recognizable. Consequently, landmark Assisted CycleGAN is proposed to define consistency loss using facial landmark features to guide the training of local discriminators in CycleGAN.

4.3. Unpaired Photo-to-Caricature Translation

Cao et al. [4] proposed a learning-based approach to solve the conversion from ordinary photographs to cartoons. A two-way model with a coarse-distinctive and a fine-distinctive discriminator is designed in order to take into account both local statistics and global structure during the conversion. For the generator, perceptual loss, adversarial loss, and consistency loss are utilized to achieve representation to learn in two different domains. Moreover, an auxiliary noise input can be used to understand the style.

It also presents a generative adversarial network (GAN) for photo-to-comic transformation without paired training datasets: the CariGAN. It uses two modules to explicitly model geometric exaggeration and appearance stylization, one is CariGeoGAN the other is CariStyGAN. In this way, a difficult cross-domain transformation problem is decomposed into two simpler tasks. Compared to advanced methods, CariGAN generates caricatures that are closer to hand-drawn while better maintaining the personality characteristics of the original face. In addition, the user is allowed to control the degree of exaggeration and variation of the shapes, or to give example caricatures to generate the corresponding styles.

4.4. Cartoon Adversarial Generation Network

Researchers at Tsinghua University have proposed CartoonGAN [5], a comic style-based generative adversarial network that can train a comic style migration model. Previous image styling algorithms based on generative adversarial networks often require two sets of corresponding images to be trained to obtain better results, such as CycleGAN, which also makes the training data difficult to obtain. The paper proposes a GAN network architecture dedicated to cartoon style migration and two simple and effective loss functions.

The cartoon generative adversarial network mainly proposes a cartoon stylized framework for generative adversarial networks, which directly trains the captured images with the cartoon images without matching them one by one and is easy to use. And the authors propose two kinds

of losses, one is semantic content loss, which is a sparse regularization constructed in the high-level feature graph of VGG network to cope with the large number of style variations between photos and cartoons; the other is an edge-promoting adversarial loss to maintain clear edges. To improve the convergence of the network to the target, an initialization phase is further introduced in this paper. The GAN framework consists of two CNNs: a generator, which is trained to produce outputs that make the discriminator indistinguishable, and a discriminator D, which classifies whether the image is from a real target or a synthetic one. The generator uses a convolutional layer for downsampling, follows a layout of eight residual blocks, and finally upsamples the image by microstrip convolution. The discriminator network D is used to determine whether the input image is a real cartoon image or not. The discriminator network has shallow layers, in fact, the discriminator network is a classification network, the discriminator mainly discriminates whether there are obvious boundary lines, the structure is two convolutional layers for downsampling, and then the convolutional layers return the classification results.

5. RESEARCH METHOD

5.1. Convolutional Neural Networks

Convolutional neural networks are the most widely used of all kinds of deep neural networks and have achieved good results in many problems of machine vision, in addition to its successful applications in natural language processing, computer graphics, and other fields. In 1989, LeCun [6] proposed a convolutional neural network that is quite efficient for handwritten character recognition, which is also the origin of many convolutional neural networks nowadays.

After nearly two decades of neural network coldness, the AlexNet network was proposed in 2012, which won the ImageNet competition at that time. the parameter scale of the AlexNet network became larger, the convolutional layers of the network became deeper, with a total of five convolutional layers, and the maximum pooling layer was added to the first, second, and fifth convolutional layers to reduce the computation, and finally, the fully connected The network divides the convolutional layers into two parallel networks, which can effectively reduce the computation and improve the computational efficiency.

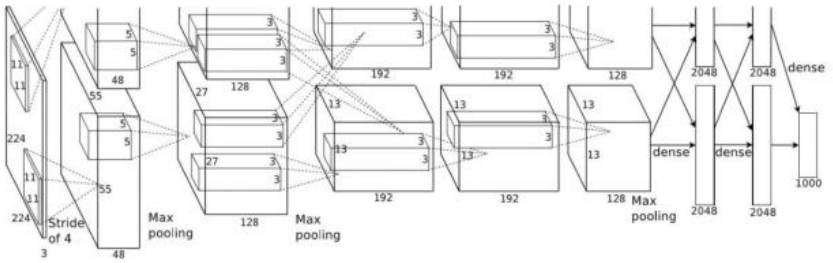

Figure 2. AlexNet Network Structure Diagram

GoogLeNet was proposed two years after AlexNet, the key in this network is the Inception block, that is, the input image is extracted with different scales of features, this mechanism change reduces the number of parameters to one twelfth of AlexNet. The network integrates multi-scale convolutional kernels and pooling layers, which effectively reduces network parameters, prevents overfitting and reduces computational effort to improve efficiency.

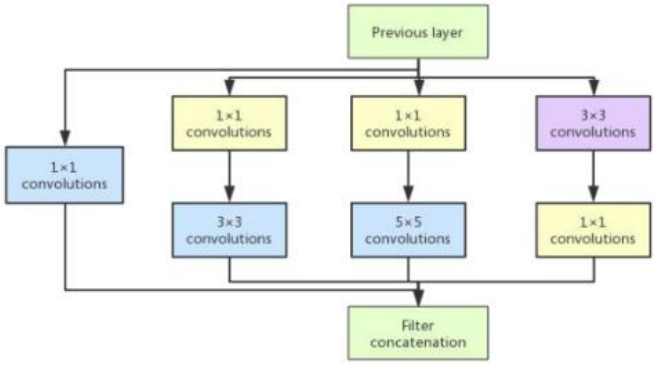

Figure 3. GoogLeNet Inception Modular

ResNet residual neural network was proposed by four Chinese including KaimingHe from Microsoft Research [9], the network structure of ResNet can pass down all the previous parameters, and the accuracy of the network is greatly improved. However, when the number of network layers is increased to 1202, the result decreases due to the overfitting caused by the deep number of network layers. The main core of the residual network is to pass the information of each layer directly to the output, which will not lead to degradation problems due to the increase of the number of network layers, and the accuracy rate can be increased based on the increase of the number of network layers to ensure the integrity of the extracted features.

5.2. CycleGAN

The field of image transfer is the domain of GAN networks, and recently many people have applied CycleGAN networks to the field of image style transfer.

Figure 4 shows structure of general GAN which is composed of generator G and discriminator D. The generator G generates data G(z) from a random input z, and makes the generated data as close to the real data as possible. On the other hand the discriminator D tries to distinguish the real data from the generated data G(z) as much as possible. The two networks are always playing a game, in which G gradually gains the upper hand and the generated data is no different from the real data.

Figure 4. General GAN

The goal is to realize the data migration of two domains, to realize the translation between images with the help of GAN, as shown in Figure 5. There should be two discriminators of domains, and each discriminator will judge separately whether the data of their respective domains are real data. As for the generator, the image translation needs to turn the image of domain A into the image of domain B. Therefore the generator is somewhat like the autoencoder structure, except that the output of the decoder is not the image of domain A, but the image of domain B. To make full use of the two discriminators, there should also be a translation back, which means there is another generator that translates the data from domain B to domain A.

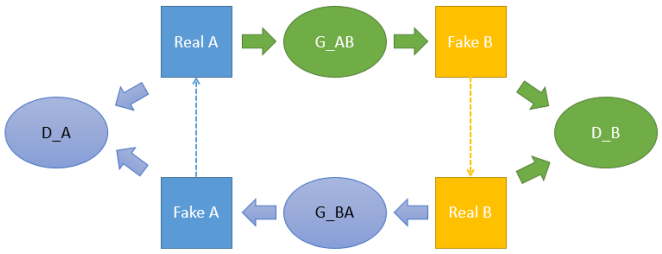

Figure 5. Data migration of two domains

The dashed arrow in Figure 5 indicates 'treat the image at the beginning of the arrow as the image at the end of the arrow and continue according to the flow chart'. It means that for Real A, the complete process is like this: $A_{real} \rightarrow B_{fake} \rightarrow A_{fake}$; for Real B, the process is like this: $B_{\text{real}} \rightarrow A_{fake} \rightarrow B_{fake}$. It can be seen that the whole process is a cycle for both domain A and domain B images, so it is called CycleGAN. The whole cycle can be seen as an autoencoder, the two generators are seen as encoder and decoder, and the two discriminators are criteria.
In general, the two generators are designed in such a way like:

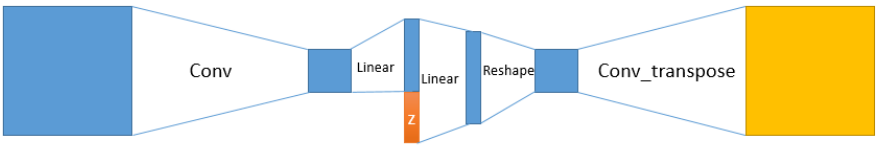

Figure 6. The two generators

Z controls some properties in G such that the generated results are not unique and can be diverse [12]. The process of CycleGAN is clarified to write its objective function as follows.

For discriminator A: $L_{D_A} = E_{x \in P_A} \log D_A(x) + E_{x \in P_{B2A}} \log(1 - D_A(x))$

For discriminator B: $L_{D_B} = E_{x \in P_B} \log D_B(x) + E_{x \in P_{A2B}} \log(1 - D_B(x))$

For generator BA: $L_{G_{BA}} = E_{x \in P_{B2A}} \log D_A(x) + \lambda E_{x \in P_A} \| x - G_{BA}(G_{AB}(x)) \|_1$

For generator AB: $L_{G_{AB}} = E_{x \in P_{A2B}} \log D_B(x) + \lambda E_{x \in P_B} \| x - G_{AB}(G_{BA}(x)) \|_1$

Adding refactoring error terms for generators, like pairwise learning, can guide the two generators to better perform the task of encoding and decoding. In turn, the two Ds serve to correct the encoding result to conform to the style of a certain domain. Not only does the structure is simple and effective but also the data of the pair is not required. Cycle consistency loss has been proposed which makes generic unpaired image-to-image translation possible. Given only two domains of image collections, CycleGAN can explore the collection-level supervised information and realize image transfer.

5.3. Generative Networks Based on Content Invariance

This paper studies the method of generating face cartoon drawings in the absence of paired experimental data. In the absence of paired data, the content features of the images need to be

constrained by indirect means, and a cyclic generative adversarial network was first proposed in the literature [14] to achieve the interconversion of the two styles, which ensures the invariance of the content by reconstructing the generated image out of the original image. The same principle is used in the literature [15] and [16], with the difference that the encoding network is divided into a style encoder and a content encoder, which encodes the image to be converted and the style image, and then performs the fusion to go through a decoding network of a specific style.

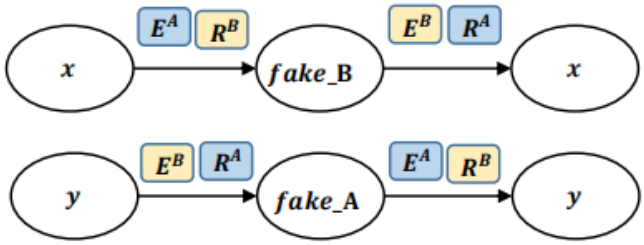

Figure 7. Network Structure Comparison [14]

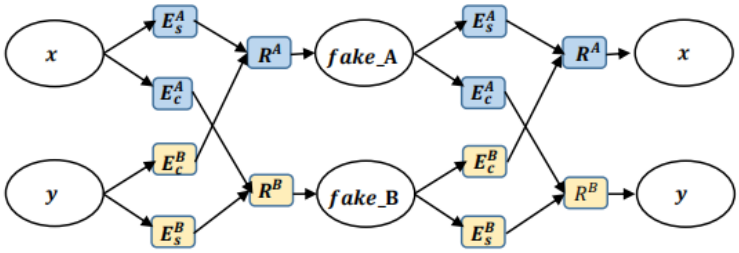

Figure 8. Network Structure Comparison [16]

The above figure shows the network structure of the literature [14] and the literature [16], and the encoders or decoders used by both methods are designed for different styles. The two styles are noted as style A and style B; x ∈ A, y ∈ B. where the network used for the different styles is distinguished by two colors in the figure, blue indicates encoding or decoding of images or features of style A, and yellow indicates encoding or decoding of images of style B. The encoder is denoted as E and the decoder is denoted as R, and the style is distinguished by superscripts A and B. The content and style encoders are distinguished by subscripts c and s.

The method in Figure (a) directly converts the input image to the corresponding style image by using different style encoders and decoders.

Since the style conversion is essentially a combination of the content information of the image to be converted and the style information of the reference image. Therefore, the method in Figure (b) extracts the content x and the y for the two input images x and y, respectively style features, and then the target images are obtained by the decoding network of corresponding styles. Both methods above restore the input image x or y by encoding and decoding the generated image, a process called image reconstruction.

In the style conversion task, the purpose of image reconstruction is to ensure the content information of the original image, while the confrontation is to make the generated image with a specific style, a certain balance needs to be maintained between the two, if the content

information is protected more, the network will choose to ignore the style information; on the contrary, the content of the generated image will not be guaranteed.

However, training the interconversion of the two styles requires facing a problem in that the network needs to build additional models. for different styles of encoding and decoding networks and discriminative networks, which means that additional memory is occupied during the training phase space in the training phase. In the testing phase, only the A to B conversion needs to be implemented, and the extra models should be minimized. This study aims to find a method that can be trained in only one direction to achieve the generation of face cartoons, so it is necessary to find another reconstruction method that can be used to guarantee the invariance of the content and at the same time be able to reach a balanced state with the styles.

Derived from the idea of the literature [16], the desired content features and style features can be extracted by constructing a content encoder and a style encoder separately, and to make the encoding network general, the same network is used for the extraction of content or style features for both style images. The two main reasons are as follows.

Firstly the so-called content features mainly include the shape structure information of the face, for the extraction of the image content features should not be affected by the image style, i.e. the content coding network should be general for all face images. Whether it is a real face image or a style image, the content features should be able to be extracted correctly.

Secondly, in the style migration task, style can be used as an attribute, and for the same content feature, based on different style attributes, it should be possible to obtain images with different styles and keep the content unchanged. Further, in the style conversion, the style features are used as the condition of conversion, and the style features are fused with the content features, and then the decoder is used to get images of different styles. Although using different style encoders for different styles of images can get more style features, it increases the complexity of the network. In deep networks, the classification of images is performed by extracting images with classification is performed by extracting discriminative features from images, and for style determination, it can also be The discriminable features are extracted from the image to determine the style of the image, and such discriminable features can be used as the style features of the image. can be used as the style feature of the image.

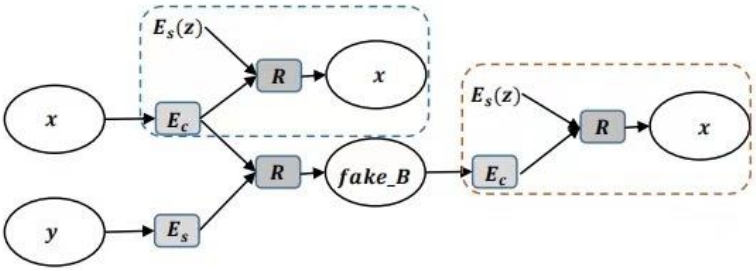

Figure 9. Generate network structure

Since the network only needs to be trained to convert from style A to style B, the reconstruction of the image also only needs to reconstruct the input image. To ensure content invariance, the content encoding of the generated image should be as similar as possible to the content encoding of the input image, so the content encoding is performed on the generated image, and then the style features of the real face are fused to recreate the original image, i.e., the orange dashed box part in the figure. The reconstructions all use the style features of the real face, and theoretically, the style features of different images of the same style should be as similar as possible. Therefore,

during the training process, for each input image to be transformed, a random image of the same style is input to extract its style features, the so-called feature input for reconstruction.

5.4. Frechet Inception Distance (FID)

Evaluation in supervised image classification is simple. It requires comparing the predicted output with the actual output. With GAN, however, some random noise is passed in to get this fake generated image. The goal is to make the generated images look as realistic as possible. Then it is necessary to accurately quantify the realism of the generated image or to accurately evaluate the GAN. It will start by setting two simple properties for the evaluation indicators: Fidelity: The GAN is expected to generate high quality images; Diversity: The GAN should generate images that are inherent in the training datasets.

Therefore, evaluation metrics should be evaluated for both attributes. But comparing fidelity and diversity at the same time can be challenging, as it is then not clear what exactly should be compared. Therefore the two widely used methods for image comparison in computer vision are: Pixel distance: This is a simple distance metric-the pixel values of two images are subtracted from each other. However, this is not a reliable metric; Feature distance: This uses a pre-trained image classification model with an intermediate layer of activation. This vector is a high-level representation of the image. Using this representation to calculate the distance metric will be more stable and reliable.

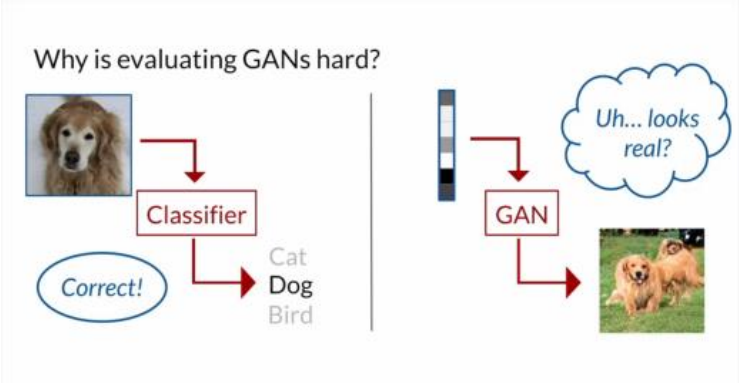

Figure 10. GAN Evaluation

This is one of the most popular metrics used to measure the distance of features between a real image and a generated image. Frechet Distance is a measure of the similarity between curves, which takes into account the position and order of points along the curve. It can also be used to measure the distance between two distributions.

The Frechet Distance is used to calculate the distance between two "multivariate" normal distributions. For a "univariate" normal distribution, the Frechet Distance is:

$$d(X,Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

where μ and σ are the mean and standard deviation of the normal distribution, X and Y are the two normal distributions. Feature distances as described above are used in computer vision and especially in GAN evaluation. An Inception V3 model pre-trained on the Imagenet datasets was used. Activations from the Inception V3 model are used to aggregate each image, which is why the score is named "Frechet Inception Distance".

The Frechet Inception Distance of the multivariate normal distribution is given by the following equation:

$$FID = \| \mu_X - \mu_Y \|^2 - Tr(\sum_X + \sum_Y - 2\sum_X \sum_Y)$$

where X and Y are the true and fake embeddings, respectively, and are assumed to be two multivariate normal distributions. μX and μY are the magnitudes of vectors X and Y.

FID score represents the distance between the feature vectors of the generated image and the feature vectors of the real image. The closer this distance is, the generated model is better, which means the image has high definition and rich in diversity.

6. EXPERIMENT

I have already trained a generative adversarial network to generate cartoon faces after showing pictures of many real faces. Most of them here are from the CycleGAN implementation in PyTorch.

I use the CelebFaces Attributes Datasets as training data, which is a large-scale face attributes datasets with more than 200K celebrity images, each with 40 attribute annotations. The images in this datasets cover large pose variations and background clutter, which means CelebA has large diversities, large quantities, and rich annotations. The other datasets is called selfie2anime. The size of this datasets is 3400, and that of the test datasets is 100. All anime face images are sized to 256 x 256 by applying a CNN-based image super-resolution algorithm.

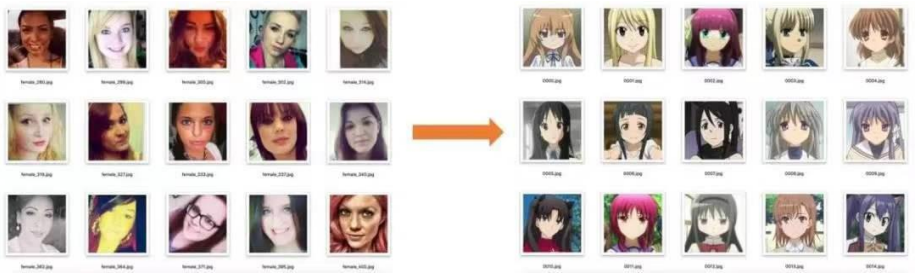

Figure 11. Datasets sample images

I have done several experiments to change the transformation effect of the images by changing the parameters, here are a few examples. In this time I changed the learning rate from 0.0002 to 0.0016:

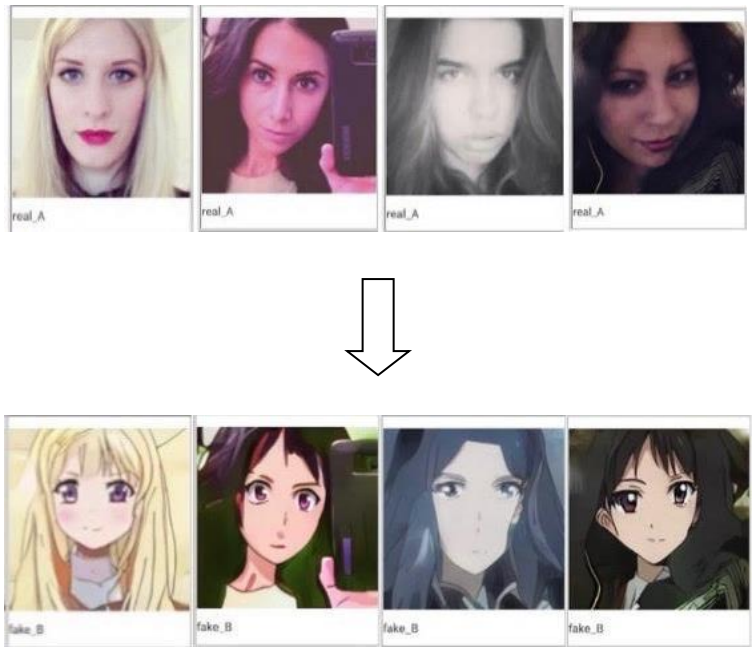

Figure 12. Experiment

7. CONCLUSIONS

The influence of cartoons in life is enormous, such as Japan's Black Deacon, Cherry Puff and other classic anime; the U.S. Disney, etc. is still being loved by countless fans around the world. Since the cartoon style is different from the oil painting style in style transfer, the cartoon needs a neat and clear border to avoid a large number of uneven color blocks.

Style transfer algorithms are increasingly studied. I have studied the common neural networks of deep learning, learned the principles and practice for the basic style transfer algorithm to lay the foundation for the next step of research. In addition, I read some papers about style transfer, studied and analyzed the cartoon translation style, to come up with a suitable model for cartoon style transfer. Based on the existing model, different parameters were experimentally analyzed to select the parameters suitable for cartoon style transfer, the datasets mentioned in this paper was used for training, which means the deep learning based cartoon style transfer algorithm was implemented. It is possible to generate cartoon style avatars with clear lines and simple character features, the results of different parameters are compared, with the aim of improving the algorithm and making the generated cartoon style avatar better.

At present, there is some progress in cartoon style transfer, but there is still a difference for real cartoon avatars, the details of the transfer are not very good, and the details of the cartoon style are still lacking. The next work will focus on the following aspects: first of all, more research on the details to achieve a more realistic cartoon style transfer effect, secondly adjust the parameters to make the characters retain more features. Hopefully, after the details are improved, it can be applied to daily life, saving more time and creating more value.

ACKNOWLEDGEMENTS

At this point in the writing, although there is reluctance, also had to say goodbye to the past. When I look back, I realize that what I want to say most is I would like to thank you. First of all, I would like to thank my supervisor, Prof. Murakami Since I entered his lab as a graduate

student, I have benefited a lot from his rigorous and serious research attitude and practical research style; I thank Prof. Murakami for his guidance and help in my research so that I can overcome one problem after another; I thank Prof. Murakami for his tireless teaching so that I always keep a positive attitude and try my best to accomplish everything.

Thank you to my labmates. During my graduate studies, my classmates provided me with a lot of help, and I learned from them the excellent qualities of aggressiveness, diligence, optimism, and perseverance. Finally, I would like to thank my family for their support, whether, in study or in life, the warmth of my family is the motivation for my efforts, and I will live actively to repay their love.

REFERENCES

[1] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge; Image Style Transfer Using Convolutional Neural Networks; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros;Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223-2232

[3] Ruizheng Wu, Xiaodong Gu, Xin Tao, Xiaoyong Shen, Yu-Wing Tai, J iaya Jia; Landmark Assisted CycleGAN for Cartoon Face Generation

[4] Kaidi Cao, Jing Liao, Lu Yuan; CariGANs:Unpaired Photo-to-Caricature Translation; ACM Transactions on Graphics, Vol. 37, No. 6, Article 244. Publication date: November 2018

[5] Chen Yang, Lai YuKun, Liu YongJin. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization [C], IEEE Computer Society Conference on Computer Vision and Pattern Recognition.Salt Lake City:IEEE Computer Society,2018:9465 - 9474

[6] Furlanetto R W, Underwood L E, Van Wyk J J, et al. Estimation of somatomedin-C levels in normals and patients with pituitary disease by radioimmunoassay[J]. Journal of Clinical Investigation,1977, 90(3):648-657.

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C], Advances in Neural Information Processing Systems, Lake Tahoe:Neural information processing systems foundation,2012:1097-1105

[8] Szegedy c, Liu W, Jia Y, et al. Going deeper with convolutions[C], IEEE Conference on Computer Vision and Pattern Recognition, Boston:IEEE Computer Society,2015:1-9

[9] He K, Zhang x, Ren S, et al. Deep Residual Learning for Image Recognition[C], IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas:IEEE Computer Society, 2016:770-778.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large- -Scale Image Recognition[J]. Computer Science, 2014.5(2):230-237.

[11] Goodfellow I J,Pouget-Abadie J,Mirza M,et al.Generative Confersarial Nets[C], International Conference on Neural Infonmation Processing Systems, 2014:2672-2680.

[12] Zhu J Y, Park T, Isola P, et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks[J]. arXiv preprint arXiv:1703.10593, 2017.

[13] Gao H, Zhuang L, Laurens M, et al. Densely Connected Convolutional Networks[C], IEEE Conference on Computer Vision and Pattern Recognition,Hawaii:Institute of Electrical and Electronics Engineers Inc.,2019:2261 - 2269

[14] J. Y. Zhu, T. Park, P. Isola, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. IEEE International Conference on Computer Vision. 2017: 2223-2232.

[15] Y. Lu, Y. W. Tai, C. K. Tang. Attribute-Guided Face Generation Using Conditional CycleGAN[C]. European Conference on Computer Vision (ECCV). 2018: 282-297.

[16] H. Kazemi, F. Taherkhani, N. M. Nasrabadi. Unsupervised Facial Geometry Learning for Sketch to Photo Synthesis[C]. 2018 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2018: 1-5.rs.