Research on the algorithm of fusion of specific text and face features to generate new images

**Abstract:** Although generative adversarial networks have achieved great success in the field of face image generation and editing, it is still a major challenge in computer vision to find the direction that can manipulate the semantic attributes of faces in its potential coding space. The realization of this challenge requires a large amount of labeled data to continuously optimize the network, and there are many difficulties in collecting and labeling similar data. Such as the high technical threshold and a large number of labor costs. Some recent work has attempted to overcome the labeled data shortage with the help of pre-trained models. Although this approach has been verified to complete the above tasks, the accuracy of the operation and the authenticity of the results can not meet the needs of real face editing scenarios. When the generative adversarial network is used for image generation, problems such as low resolution, fuzzy edge and loss of identity information feature are easy to occur. Based on this, the algorithm of fusion of specific text and face features to generate new images is deeply studied in this paper. By analyzing the correlation between text description and face features, a novel algorithm is proposed, which can effectively fuse text description with face features to generate high quality and personalized new images. The experimental results show that the experiment using the pre-processed CelebA dataset shows that the algorithm has a significant improvement in subjective visual effect, and a stable improvement in PSNR and SSIM compared with the existing methods. The algorithm has achieved a significant improvement in the accuracy and diversity of image generation. It lays a good foundation for further research and application of the technology of fusion of text and face features to generate new images. This research is of great significance for promoting the development of artificial intelligence field.

**Key words:** facial features; Generate adversarial network; Face image generation

# 1 Introduction

For the researchers of face image generation, this is a challenging and meaningful topic. Along with advances in style-based generation neural networks (stylegan) and related technologies, a new mode of photo generation has succeeded in creating high-quality photography. In order to improve the ease of use of this process, some scholars have proposed a series of ideas for quickly completing the shooting and post-production work according to specific parameters, such as character sketches or text descriptions. The most intriguing of these ideas are those that combine machine learning and human-computer dialogue to facilitate the translation of spoken words into actual images - an approach that relies on deep learning to understand text information provided by users and then use it to guide artificial intelligence systems to create corresponding images; Compared with traditional methods, this strategy does not need to introduce complicated optimization problems in multi-dimensional space, nor does it need to use any probability theory tools or other simulation techniques to effectively reduce the difficulty and improve the overall work efficiency. However, it should be noted that although people can provide their needs through various ways, such as natural language and images, the results they expect are often different due to various factors, so how to accurately and efficiently meet people's wishes has become one of the urgent problems to be solved at present

The current text-driven facial image production and modification technology mainly relies on multi-level deep learning system to achieve, which will integrate the low-pixel level photos generated by each link to get the final graphics. The problem with this structure, however, is that the training process is extremely unreliable and the resulting image looks like a patchwork of various character elements rather than a realistic work; Especially when it comes to creating difficult visuals based on written material, the authentication function is not able to provide adequate guidance. In detail, this strategy generally consists of three steps: Each step consists of a group of producers and evaluators (i.e., two neurons) and runs in a synchronized state until three different sizes such as 64×64, 128×128 and 128 are completed Three works of different sizes of 256×256, etc. - First, the original basic painting with outline but no color will be gradually improved into a fine layout photographic film that meets the requirements of word expression. However, the commonly used pattern recognition tools in such schemes cannot be applied to the leap-forward information exchange, which results in the failure of feature matching when creating novel images based on literature or the unexpected behavior of irrelevant parts. Therefore, we used the previously optimized StyleGAN2 machine and the human-machine interactive CLIP algorithm to build a new basic architecture. Comparatively speaking, compared with the previous version, the latest stylized artificial object 2 can better solve some characteristics of the flaws and improve the overall freshness and clarity to become one of the most efficient optimal solutions today. CLIP is a pre-trained model that includes image encoder and text encoder. By training 400 million image-text pairs, The ability to accurately measure semantic similarity between input images and text descriptions.

In view of the technical difficulties of text-driven image generation and editing, the key problems include the following :(1) The natural speech parsing ability still faces many obstacles. Although natural language analysis in machine learning has made significant progress in many scenarios in recent years, natural language texts often contain multiple interpretations or misleading information, so it is still very difficult to fully grasp and accurately interpret this complex discourse. On the one hand, any kind of words can be combined into short sentences or even long essays, and on the other hand, similar meanings can be presented through different expressions. On the contrary, the same meaning can be expressed in various forms in English articles and other types of expressions. To solve this contradiction, we must rely on a great deal of human intelligence for logical judgment and speculation. There's also the question of how to integrate this information more effectively into the system and how to use it efficiently so as to eliminate the possibility of confusion and that's a very large and very difficult study; (2) The skills of photo synthesis need to be further improved. This is one of the central themes of computer imaging science and other related fields, and has been successfully applied to everything from typeface generation to cross-modal conversion of photography to fixing imperfections such as removing impurities and adjusting brightness. But the biggest problem at this stage is how to ensure that the resulting images are realistic while maintaining enough differentiation to meet the needs of users. Although generative adversarial networks have improved the realism of image generation in recent years, they are faced with problems such as uncertainty in training process and low quality output samples. The lack of perfect standard evaluation methods and the incomprehensible characteristics are also one of the main problems in this field. Consider that stylized Gans can effectively decompose high-level abstract features (such as human posture and personality or noise or hair strands in the resulting picture, etc.) without any guidance, and can achieve certain manipulative effects to adjust the photo composition. Therefore, how to use deep learning language with rich connotation and good segmentation ability to deal with facial image transformation is still a scientific and technological task to be solved. The current methods are generally to insert text expression vectors into the conditional generation adversarial network to monitor the process of image generation, such as word embedding and sentence pattern embedding to perform language level supervision, and identify meaning according to the differences of embedding vectors. However, no effective connection points have been found in the generative adversarial network to match text embedding vectors with visual attributes. As a result, text-driven face image generation cannot guarantee the consistency of visual meaning, especially when a multi-group generative adversarial network is used to build a model, the image quality largely depends on the training stability of the generative adversarial network, but this method does not significantly improve the efficiency of image generation. Therefore, solving the image-text semantic unification problem and maintaining the network training to generate high-quality images have become the main problems in the production and modification of text-guided images at present.

# 2 Related Work

## 2.1 Image-text joint mode

To achieve the goal of text-driven image generation and editing, the first condition is to match the visual features with the relevant text information. The current main strategy is to obtain clear word-level training responses through discriminators, while achieving synchronization between vision and meaning through a series of image-text similarity assessments. The aim is to build a strong relationship between text and image with the help of matching associations, which in turn conveys guidance signals to the generator. Based on granularity of expression, most technologies can be grouped into two broad categories: those that use deep neural networks to capture the global characteristics of the two and then compare them; The second is to perform example-based image text matching and learn to recognize the relationship between words and image parts.

For Category 1 research, problems such as using language to describe image search, generate image captions, or solve visual problems are all based on the combination of language and image. After BERT achieved significant results in multi-language processing tasks, more and more technologies began to use Transformer to build federated representations. However, thanks to the advent of Contrast Language Image Pre-Training (CLIP), we have entered a new and broad field of multimodal potential coding that can be directly applied to the computation of semantic associations between words and images. CLIP is proposed by OpenAI, and its training sample contains more than 400 million image-text combinations, which can synchronize graphics and text encoding to achieve cross-domain language image joint representation. CLIP performs particularly well on multiple modal tasks, such as image finding, geolocation, video behavior recognition, and more, and only requires unsupervised learning to achieve results similar to those of mainstream supervised algorithms. Although CLIP's design concept seems simple, the excellent results achieved with this simple strategy further confirms the huge growth potential of multimodal models.

The second type of research focuses on the understanding of the relationships between important elements, and makes significant breakthroughs in this field by learning the associations that co-occur with objects to construct the corresponding relationships. However, such studies only focus on the simple connection between key elements, because they mainly rely on the relationship between prominent elements, thus ignoring the role of other secondary elements in this process and their states, especially the lack of attention to environmental information, which makes it difficult for us to achieve accurate language-visual combination representation.

## 2.2 Face feature generation

Similar to the method of facial feature generation and appearance style transformation, both need to try to maintain the information of the initial personal identity while changing some of its main characteristics. The traditional facial feature generation method relies on image processing skills, which takes out the corresponding patterns according to the desired feature style, and then pairs or replaces them into new face photos. In the field of image and video processing, scholars use the characteristics of brushes that imitate a certain painting style to simulate the artistic elements of human face area with the help of real visual effects. In the research field of image filtering, researchers use the smoothing weight function of the Sanyuan filter instead of the block region, and construct classification rules considering the directional weight function, which can not only make the image smoother but also save valuable boundary information in the image, so as to improve the results generated by image features. The Variable Auto Decoder Machine (VADM) proposed by Kindma et al. adjusts the decoder by introducing prior knowledge to the hidden unit, so that the hidden unit can be extracted or put into it to realize the face image. However, because the learning goal of VADM is pixel-based Gaussian likelihood function, the images produced by this method are often too smooth or even fuzzy. This is good for smoothing things like clouds in the sky, but it's not ideal for images of faces. Larsen et al. therefore improved VADM by adding a GAN-derived recognizer and demonstrated that this approach produced more realistic facial images. Training the variational autoencoder and the counter loss function simultaneously can avoid excessive smoothing, but may cause distortion. This method needs to consume a lot of time and labor cost to carry out the initial sampling work, and may face difficulties in practical application.

At present, the generated adversarial network is widely used to produce more realistic face images. The technique has been proven by a large number of researchers to produce high-quality images and has shown excellent results in areas such as text generation. In general, Gans employ a dynamic competitive strategy, using generators to continuously interact with discriminators to create photos that they believe to be real. However, because most GAN-based methods are limited by the number and quality of initial samples, they cannot ensure the preservation of the resulting image quality and character characteristics. Therefore, we need to explore how to add a variety of attention modes in the generative adversarial network to enhance the emphasis on local correlation, so as to optimize the processing of generative features. In addition, we also try to use multi-scale feature maps to construct gradient transmission paths to overcome the instability of generating adversarial networks and improve the model's ability to cope with data sets of various categories, sizes, and image quality. At present, comparative learning has also been integrated into generative adversarial networks to ensure that models can achieve more generative options while maintaining a high level of image quality.

In addition, the evaluation of the results of this type of work needs to compare its numerical performance with the degree of variation, and it also needs to take into account the individual perceived experience of the practitioner. A more impartial way to combine the above two factors is to have practitioners participate in the assessment and rate the generated facial images.

## 2.3 Generate adversarial networks

As shown in Figure 1, Generative Adversarial Networks (Gans) are a deep learning model of game theory that consists of two main components, a generator and a discriminator. The two components are continuously trained interactively through adversarial learning, thus achieving a competitive and self-optimizing process to improve the performance of the entire model. The generator G is a network responsible for generating images, it receives a random noise z, through which it generates a false image similar to the real data, and records the resulting picture as G(z). The discriminator D is responsible for determining whether the input image is a real image or a composite image generated by G. The core idea of GAN is to make the generator generate more realistic false data through adversarial learning, and make the discriminator improve its discriminating ability, so as to achieve the purpose of generating false data similar to the real data. In an ideal state, the generator G can generate a very realistic image G(z), making it difficult for the discriminator D to judge the truth or falsity of the generated image. At this time, D(G(z)) is close to 0.5, and the generator G is considered optimal and can successfully generate high-quality images.

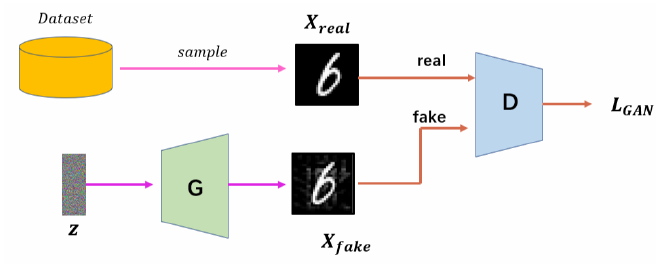


Figure 1 Generates the adversarial network structure

The training of GAN network can be regarded as a process of common progress of the two models. In the training of GAN, "progress" means that the samples generated by the generator are gradually close to the real samples, and the discriminator's judgment accuracy of the true and false samples is also constantly improving. This process is mutually antagonistic and promoted, and the generator and discriminator improve each other's performance through continuous training and adjustment to achieve the final optimization goal. Through this way of confrontation and game, the generator and discriminator continue to optimize, and the generated samples are more and more close to the real samples, so as to achieve the function of generating models. However, since the generator learns the feature distribution of the entire dataset sample, the generated result is also completely random, and it cannot be specified for the generation of specific content. In order to solve the defects of the original GAN networks, conditional adversarial generation networks (cGANs) are proposed. Unlike the basic GAN model, cGAN can guide the network to produce relevant results by specifying label data. In order to achieve this, cGAN needs to concatenate the input random noise information with the label category, and the generator takes the concatenated result as input to generate the data that meets the requirements. The discriminator also needs to concatenate real or generated data with the corresponding label category, and then input it into its neural network for recognition and judgment. In this way, cGAN can introduce label data into the network at the same time as it generates data, producing more meaningful results.

# 3 Image generation based on specific text and face features

## 3.1 New generators supported by dual input sources

In order to solve the problem of the key features of the local face, the whole input picture is divided into two main parts: one is the area that needs to maintain the personal identification characteristics as much as possible; The other is the purpose constraint section generated based on the paired sample. In contrast to the hybrid processing of input pictures and target conditions in other technologies, our generator uses separate branches to perform the task of encoding the entire input picture and key features. These two separate paths are finally connected through a series of style integration modules to achieve the combination of comprehensive features and important information texture features, and then the decoder is used to create a new photo.

(1) Dual input source feature coding

In the initial stage, the whole input image Iin is processed, and the global feature decoder constructed by 28 low level extraction convolution layers is used, which accords with the general encoder setup principle. Next, we will focus on the local important features of the input image Is. The first step is to use feature parsing and decoding to separate it into style and deep spatial information. This process needs to start from the semantic mapping of the image, and then use the binary mask component Mi corresponding to the position of Iin to obtain a variety of decomposition features Isi. Specifically, you can refer to the description in formula (1), where the "·" represents the dot product between the corresponding elements.

（1）

The resulting is then input into the texture encoder to obtain the key feature style encoding , as shown in formula (2). All encodings are then superimposed to obtain the texture feature Cs that needs to be progressively changed in the model as a whole.

（2）

In texture encoder, in order to better adapt to diverse and changing task requirements, we use a pre-trained VGG network to enhance the generalization performance of the encoder. The specific internal process is shown in Figure 2.

These strategies encode the process by carefully partitioning and disassembling key elements, which helps prevent the model from processing all the complex features of the entire face area at once, but instead coding them individually in each specific cell domain, significantly increasing the model's convergence rate and resulting in a more realistic visual representation in a shorter time. This design benefits from pre-training, it has a strong feature extraction ability, especially for small sample size scenes, and the resulting texture features are symbolic of the abstract features generated after deep convolution, they reflect the detailed features of the face.

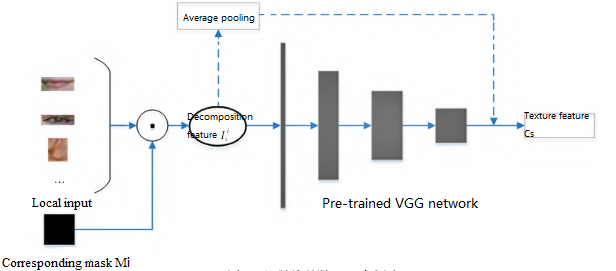


FIG. 2 Schematic diagram of texture

(2) Texture feature fusion and image construction

Texture feature fusion aims to incorporate texture features obtained from locally important features into the overall features generated through all image input, so as to reflect the overall individual characteristics that must be changed. This step consists of a series of concatenated style mixing units F, each containing AdaIN and residual convolution blocks. For each level of connected blocks, its input is the feature Ft-1 and texture feature Cs output by the previous layer, and then Ft-1 and Cs are combined with each other in the residual convolution block to perform operations, and the final result is summed with Ft-1. For the specific operation process, please refer to formula (3).

（3）

Where φ represents A series of operations carried out in the cascade block, and parameter A represents the affine transformation parameters that AdaIN layer needs to learn, including scaling parameters and displacement parameters, which can be used to normalize the features of each layer. Finally, the output F from the last cascade block is fed into the decoder, which follows the normal decoder configuration and builds the final image I\_out through multiple deconvolution layers.

## 3.2 Multi-level semantic information separation based on StyleGAN2 model

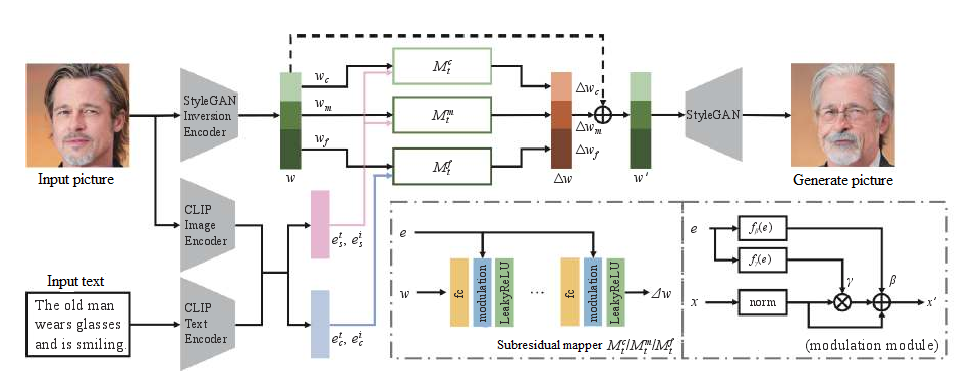


Figure 3 Network structure diagram

As evidenced from StyleCLIP, each level produces, at a different level, a specific level of information content on the picture generated by Stylized GAN2; The deeper the depth of the same level, the higher the level of specific information it produces. Therefore, in order to achieve the purpose of more accurate control of facial photo synthesis and correction, we introduced a new concept in the paper -- multi-level residual projection Mt to calculate the corresponding balance amount according to the target value of the text description for various degrees of specific facial features. This number can be directly used to adjust the existing image heritage code to complete the creation or modification of specific facial features. Different from the mapper proposed in StyleCLIP, the subresidual mapper proposed in this paper can map the image and text encoding respectively to the latent space of StyleGAN2 according to the newly defined semantic information level (face attribute and attribute state), and predict the residuals corresponding to the latent encoding while realizing attribute decoupling.

The mapper contains three sub-residual mapper 、and , which correspond to high, medium and low level semantic information respectively. Firstly, StyleGAN2 inversion model is used to encode the input image. According to the semantic information level, the different semantic information of the inversion potential encoding w=() is input into the corresponding sub-residual mapper 、and respectively, where 、and correspond to the high-level semantic information, intermediate semantic information and low semantic information encoded by the input image inversion. Then, the potential face attribute information ={ } from the image and text content obtained by CLIP image and text encoder is input as a condition the sub-residual mapper and corresponding to the high and middle level semantic information. The attribute state ={ } from the image and text content is input as a condition to the sub-residual mapper corresponding to low-level semantic information, then the internal composition of the multi-level residual mapper can be expressed as:

（4）

Under certain conditions, maintaining the existence of a language level and the stability of the relevant level of language information can also perform various tasks, so our proposed approach can also allow only any 1 or 2 of the 3 mappers to be trained to achieve different granularity of face processing effects. In the process of training, we will continue to improve the residual estimation ability between multiple mappers, so that the output is more close to the text description while maintaining and retaining the independent face attribute.

## 3.3 Feature fusion based on conditional modulation module

In order to reasonably integrate the potential encoding from StyleGAN2 inversion model and CLIP model, a new modulation module is designed in this paper. The function of this module is to maintain the identity information of randomly generated image or input image itself when the image resolution is low in the generation process. Fill in more detail at higher resolutions. As shown in Figure 4, each subresidual mapper consists of 5 subblocks, each of which consists of a fully connected layer, a modulation module, and a nonlinear activation layer. The newly designed modulation module does not simply superimpose the converted CLIP latent code and the inverse image latent code for feature fusion, but uses the conditional latent code to modulate the intermediate output of the previous layer. Therefore, the modulation module can be formulated as:

（5）

Where, and are the mean and standard deviation of x, respectively. and are fully connected networks (consisting of two fully connected layers, an intermediate normalized layer and a ReLU activation layer). Inspired by the work on conditional image translation, if no conditional input is provided for a face attribute or attribute state, the case is represented as or. In this way, we have the flexibility to allow users to edit only face attributes, only attribute states, or both face attributes and attribute states.

# 4. Experimental results and analysis

## 4.1 Data set preprocessing and experimental environment

Since the CUB-200-2011 and MS-COCO data sets themselves do not include layout maps for training, data sets need to be pre-processed before training on these two data sets. CUB-200-2011 data set directly extracts the bounding box of each image, fills the region as a real layout map. There is a problem that the category of annotation information of MS-COCO does not match the text description. For example, when the image is described as "afamily", the annotation of the image is plural "person". Direct use of annotation data may lead to the failure of the layout generation network to learn the correct layout information. Therefore, on the MS-COCO dataset, it is necessary to use the GLIP model to re-label each image to make the category of the label conform to the description of the text. The server operating system used to train the algorithm model in this paper is Windows, the training GPU is TitanX, the CPU is Intel Corei7-9700K, and the memory size is 64GB.

In view of the low quality of the original image sample, the training data sample must be preprocessed first. In this study, three preprocessing strategies, namely mean fuzzy, median fuzzy and custom fuzzy, are used to add fuzzy noise to the samples. Specifically, for mean fuzzy preprocessing, we randomly set the fuzzy range in the range of 9 to 15 pixels; For the median fuzzy preprocessing, the fuzzy range is set in the range of 7 to 15 pixels. As for custom ambiguity, the noise is calculated and added by using a single-channel floating-point matrix. These blurring effects can be seen in Figure 4.

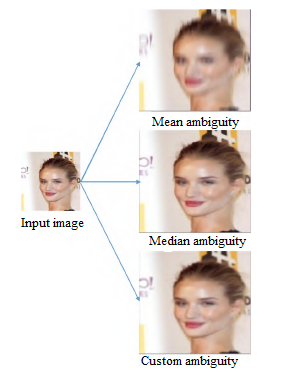


FIG. 4 Effect diagram of fuzzy processing

## 4.2 Experimental process and evaluation index

In this paper, a variety of existing algorithm models such as CycleGAN are selected as reference objects, and data sets containing three types of different noises are used as initial data sources to train these reference models and the optimized model in this paper, and then actual tests are conducted to compare the experimental results. In this phase, two controlled experiments will be performed: one is a portrait with a lens generated from a portrait without a lens; The second is the reverse operation.

The results of two rounds of experiments are compared and analyzed. The results are evaluated subjectively, and the feasibility of the model optimization is verified quantitatively by using pixel technology. Peak signal-to-noise ratio (PSNR) is a common and effective quality evaluation method for image pixel statistics. It evaluates the quality of the target image statistically by measuring the gray difference between the target image and the corresponding pixel of the reference image. The calculation formula of PSNR index is shown in equation (6) :

（6）

Where MSE represents the mean square error of the original image and the processed image. Calculating the variance of each pixel value of two images in the same position is often used to compare the degree of distortion between the newly generated image and the original image. The larger the PSNR index value, the higher the quality of the generated image. When we compare the mean square error with the Q value of the image, we get the concept of PSNR. It is a key image quality evaluation criterion proposed by the communication theory. It mainly focuses on the proportional relationship between the brightest pixels and the degree of noise change. The higher the value, the more similar the generated image is to the original photo.

In 2002, Wang proposed a new image analysis method that combines the research results of image processing, image compression, and image visual quality assessment. He introduced a new concept called "structural information" and pointed out that the core task of the human visual system is to extract structural information from the background, and it is able to accomplish this task through its powerful adaptive ability. Therefore, the measurement of the degree of distortion of the image structure should be regarded as the closest estimate of the perceived quality of the image. On this basis, structure similarity (SSIM), a subjective evaluation index of artificial image quality satisfying human visual characteristics, came into being.

SSIM measures the similarity of two images from three aspects: brightness, contrast and structure. Therefore, the definition of SSIM index is shown in Formula (7) :

（7）

Generally, α=β=γ=1, SSIM value range is [0,1], the larger the value, the higher the image quality.

In order to achieve the goal of this algorithmic model, which is to maximize the reproduction of real facial features, we must take into account the observer's personal perspective and feelings. In general, the human eye is more sensitive to large changes in color and is also limited by the individual's focus of interest. This is based on the macro perspective to think about the problem, the main concern is the change of the whole picture. However, PSNR and SSIM are measured by studying the change of each pixel unit in the image. They can reveal the total transformation by the change of the number of pixel units, and make objective evaluation of human visual experience by digital means. Therefore, in the actual operation, we should flexibly use these two analysis strategies in order to fully understand the experimental results.

## 4.3 Result Analysis

First, using CycleGAN as an example, Figure 5 shows an example of the experimental results generated based on the experimental steps described.

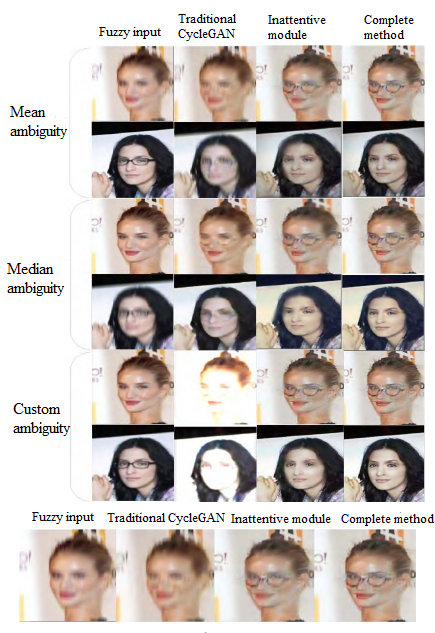


FIG. 5 Experimental effect diagram

As shown in Figure 5, it can be observed that when image generation is done directly without optimization, the resulting image quality is not good. When the input image contains noise blur interference, the traditional CycleGAN algorithm will be difficult to learn the features of the facial features, which will lead to poor quality of the final generated image, the loss of key details, and the effects of the facial features and facial edges are blurred. By using the generator structure of the algorithm in this paper and learning with dual input source coding, thanks to the independent input of local key images, a good generation of key feature areas can be achieved without adding attention guidance, but the edge outline of the central facial organ is still not clear, and identity features such as pupil color are not obvious. Therefore, after adding the attention module for continuous guidance, the final key features generated are the clearest compared with other methods, the overall visual effect of the face is better, and the features to be achieved with glasses are the most real and obvious, while there is no huge feature change. After testing different models, evaluation indexes PSNR and SSIM were calculated, and the results were shown in Figure 6 and Figure 7. Experiment A refers to generating a face image with glasses into a face image without glasses, and experiment B is the opposite.

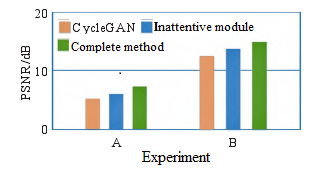


Figure 6 PSNR result diagram

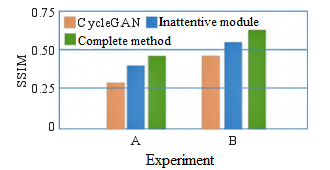


Figure 7 SSIM result diagram

PSNR is the most commonly used and widely adopted subjective measurement tool for assessing image quality. By observing the data table, we can find that the performance of the optimized model is significantly better than that of the non-optimized CycleGAN model. For the fully optimized model, the PSNR values are 7.84 and 14.66, respectively. When the attention model was used, their PSNR values were 5.46 and 13.22, respectively. As for CycleGAN, its PSNR is between 4.29 and 11.96.

In addition to the quantitative analysis of each pixel in the image, we also take into account the evaluation of the consistency of the construction between the images before and after using the SSIM. According to the results of the experimental data, the method proposed in this study can achieve the highest score standard of 0.48/0.65 SSIM; The part without attention achieved the lowest score of 0.40/0.54; Cycle GAN's performance on this task ranked second with a score of 0.33/0.47 (see Table 1).

Table 1 Experimental data

|  |  |  |
| --- | --- | --- |
| Method | PSNR/dB | SSIM |
| pix2pix | 8.76/12.75 | 0.45/0.57 |
| W GAN | 15.33/16.42 | 0.38/0.50 |
| StyIeGAN | 12.95/15.84 | 0.42/0.48 |

In summary, through the study of the experimental results, it is found that the model in this study shows excellent visual expression and can better fit the real face photos, and also shows superior performance in the quantitative comparison. For example, it has a PSNR and SSIM improvement of up to 3.85% and 38.3% over CycleGAN. In addition, its advantages are also reflected in the improvement of 2.63% and 20.4% compared with the model without the feature processing module. According to the data in Figure 6, Figure 7 and Table 1, it can be clearly seen that the image generated by the algorithm model in this paper is superior to other algorithm models, and has a higher similarity to the original image, while reducing the possibility of distortion and stable structure. These advantages can be explained as follows: First, different from the traditional learning method which uses the whole image as a whole input, we introduce the concept of hierarchical learning into the model design, so that the model can be trained separately for the key features of each part; Secondly, using the guidance mechanism of attention, continuously producing new samples helps to focus on key points to prevent too much useless information from affecting other parts of the change; Finally, the function of multiple loss functions ensures that the model maintains the initial face recognition characteristics as much as possible, and the independent discrimination after multiple cuts aims to reduce the possibility of mutual interference between various features. In general, the method in this study improves the model's recognition of important features, especially in the environment of low data quality, and strengthens the model's nonlinear transformation function. The improvement of the generalization power of the model not only contributes to the increase of the complexity of the whole model and the optimization of the operation efficiency, but also enables us to successfully extract important local information from the face image.

# Conclusion

In this paper, an improved generative adversarial network (GAN) based technology for face image synthesis is proposed. This innovative framework allows us to focus on the key elements of the face by separating and integrating some important attributes, while adding the application of attention, which in turn guides the information about the identity and maintains this information during the generation process and produces detailed characteristics. Compared with previous methods, the proposed method has better performance both in the visual perception of generated images and in the quantification of PSNR and SSIM data. However, the generation effect of screen details such as contour edges and key features can still be optimized. In further research, we can try to explore how to improve image quality and generated feature quality while maximizing the retention of identity information.

# Reference

[1] Zhang Mingjin, Wang Nannan, Li Yunsong, et al. Deep latent low-rank representation for face sketch synthesis[J] //IEEE Transactions on Neural Networks and Learning Systems, Los Alamitos: IEEE Computer Society Press, 2019, 30(10): 3109-3123

[2] Zhu Mingrui, Wang Nannan, Gao Xinbo, et al. Deep graphical feature learning for face sketch synthesis[C] //Proceedings of the 26th International Joint Conference on Artificial Intelligence, San Mateo, Morgan Kaufmann, 2017: 3574-3580

[3] Zhu Mingrui, Wang Nannan. A simple and fast method for face sketch synthesis[C] //Proceedings of the International Conference on Internet Multimedia Computing and Service, New York: ACM Press, 2016: 168-171

[4] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[C] //Proceedings of the IEEE conference on computer vision and pattern recognition, Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440.

[5] Li Yuhang, Chen Xuejin, Yang Binxin. DeepFacePencil: Creating Face Images from Freehand Sketches[C] //Proceedings of the 28th ACM International Conference on Multimedia, New York: ACM Press, 2020: 991-999

[6] Chen ShuYu, Su WanChao, Gao Lin, et al. DeepFaceDrawing: Deep Generation of Face Images from Sketches[J] //ACM Transactions on Graphics, New York: ACM Press, 2020, 39(4):72:1-72:16.

[7] Yu Jun, Xu Xingxin, Gao Fei, et al. Toward Realistic Face Photo-Sketch Synthesis via Composition-Aided GANs[J] //IEEE transactions on cybernetics, Los Alamitos: IEEE Computer Society Press, 2020, 51(9): 4350-4362

[8] Qi Xingqun, Sun Muyi, Wang Weining, et al. Face Sketch Synthesis via Semantic-Driven Generative Adversarial Network[C]//Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Los Alamitos: IEEE Computer Society Press, 2021: 1-8

[9] Yi Ran. Liu Yong-Jin, Lai Yu-Kun, et al. APDrawingGAN:Generating Artistic Portrait Drawings from Face Photos with Hierarchical GANs [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),Los Alamitos: IEEE Computer Society Press, 2019: 10743-10752

[10]Gu GH, Cao YY, Li G, Zhao Y. Image hierarchical classification based on semantic label generation and partial order structure. Ruan Jian Xue Bao/Journal of Software, 2020, 31(2): 531–543 (in Chinese with English abstract).

[11]Nam S, Kim Y, Kim SJ. Text-adaptive generative adversarial networks: Manipulating images with natural language. In: Proc. of the 32nd Int’l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 42–51.

[12]Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE,2018. 1316–1324.

[13]Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021,32(2): 327–348 (in Chinese with English abstract).