International Journal of Software and Informatics, ISSN 1673-7288
http://www.ijsi.org, ijsi@iscas.ac.cn, +86-10-62661048
IJSI, 2022, 12(2): 245–258, doi: 10.21655/ijsi.1673-7288.00267


International Journal of Software and Informatics

Research Article

# Image Style Transfering Based on StarGAN and Class Encoder

Xinzheng Xu (许新征)[1,2], Jianying Chang (常建英)[1], Shifei Ding (丁世飞)[1,2]

[1] (School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)

[2] (Engineering Research Center of Mine Digitalization, Ministry of Education, Xuzhou 221116, China)

Corresponding author: Xinzheng Xu, xuxinzh@163.com

**Abstract** The image style transfer technology has been integrated into people's lives and is widely used in practical scenarios such as artistic images, photo to cartoon, image coloring, filter processing, and occlusion removal, which bears important research significance and application value. StarGAN is a generative adversarial network framework used in recent years for multi-domain image style transfer, which extracts features through simple down-sampling and then generates images through up-sampling. However, the background color information and detailed features of characters' faces in the generated images are greatly different from those in the input images. In this paper, the network structure of StarGAN is improved, and a UE-StarGAN model for image style transfer is proposed by introducing U-Net and edge-promoting adversarial loss function. At the same time, the class encoder is introduced into the generator of the UE-StarGAN model, and an image style transfer model fusing class encoder based on a small sample size is designed to realize the image style transfer with a small sample size. The experimental results reveal that the model can extract more detailed features and has some advantages in the case of a small sample size. The images obtained by applying the image style transfer based on the proposed model are improved in both qualitative and quantitative analyses, which verifies the effectiveness of the proposed model.

**Keywords** image style transfer; generative adversarial network; StarGAN; U-Net; class encoder

**Citation** Xu XZ, Chang JY, Ding SF. Image style transfering based on StarGAN and class encoder. *International Journal of Software and Informatics*, 2022, 12(2): 245–258. http://www.ijsi.org/1673-7288/267.htm

In recent years, Artificial Intelligence (AI)[1, 2], as an important research direction in the field of brain-inspired intelligent computing, has made huge advancements. Various network structures derived from the Convolutional Neural Network (CNN) have been proposed and attracted extensive attention from experts and scholars in China and other countries. CNN has gained widespread use in many fields such as computer vision[3], natural language processing[4], speech recognition, information retrieval, recommender systems, and multimedia, which have

This is the English version of Chinese article "基于 StarGAN 和类别编码器的图像风格转换. 软件学报, 2022, 33(4): 1516–1526. doi: 10.13328/j.cnki.jos.006482"

Funding items: National Natural Science Foundation of China (61976217, 61976216)

Received 2021-06-01; Revised 2021-07-16; Accepted 2021-08-07; IJSI published online 2022-06-25

set off a wave of neural network research in industrial and academic circles and promoted the development of AI.

In deep learning, image style transfer was first achieved by CNN[5], but the image transfer result was not satisfactory due to the high requirement for training samples and slow training speed. In 2014, the Generative Adversarial Network (GAN)[6] proposed by Goodfellow *et al*. has received extensive attention for its powerful data generation ability and become an important research achievement in the field of AI. In particular, GAN demonstrates remarkable performance and great application prospects in aspects including image resolution[7–9], image compression, image style transfer, text-to-image generation[10], visual computing, and speech and language processing, and thus it is one of the research hotspots in computer vision and image processing. For image style transfer, the unsupervised training of GAN can be completed with a small dataset. GAN adopts a supervised learning approach to perform unsupervised learning tasks and uses a discriminator to supervise the learning, and eventually, it relies on a generator to learn and obtain the real data distribution or the predicted density for new image generation. As increasing attention has been paid to GAN in image style transfer which is currently the optimal network architecture among generative models, a series of excellent image style transfer models are derived, such as StyleGAN[11], Pix2Pix[12], CycleGAN[13], DiscoGAN[14], and DualGAN[15]. A GAN comprises a generator and a discriminator for learning the probability distribution of real sample data, both of which are trained under the idea of the adversarial game. The generator captures the potential distribution of real data samples from the input noise and strives to generate fake images that might be taken as real ones by the discriminator, while the discriminator does its best to discriminate the authenticity of the input images. By training, their respective generative and discriminative abilities are continuously optimized and improved until the Nash equilibrium[16] is achieved, namely when the discriminator is unable to discriminate the fake images generated by the generator.

In practical production and life, image style transfer has wide applications, such as cell phone filters, artistic images, cartoon and animation production[17], online fitting and shopping, makeup trial and removal, occlusion removal[18], and sample dataset augmentation. The study of human facial expressions[19] is of great importance in computer vision and cognitive science and has been widely applied in fields including entertainment, social contact, and face recognition. However, the expression capability of deep neural networks is constrained by insufficient facial expression datasets, which leads to low accuracy and unnoticed local details of images in the actual model training process. Considering this, image style transfer is used to augment datasets. In underground coal mines, due to poor lighting conditions and heavy dust, images captured by surveillance video devices are often blurry and dim and have unclear details. With image style transfer, the image resolution and brightness can be enhanced for mine safety and convenience. These applications of image style transfer promote scientific and technological development and people's living standards as well as crucially and greatly reduce the costs of human, material, and financial resources. The current work on image style transfer has made leaps and bounds and achieved good results. However, there are still some shortcomings and deficiencies with great room for exploration. Therefore, the use of GAN for image style transfer has a great research value.

The main contributions of this paper are as follows:

(1) in this paper, the UE-StarGAN model is proposed for image style transfer, which is based on the StarGAN model and integrates the skip connection in the U-Net network and the edge-promoting adversarial loss function.

(2) the class encoder is introduced to UE-StarGAN, and we propose an image style transfer model CUE-StarGAN fusing class encoder for a small sample size. Then, the Mish activation

function is introduced to further optimize the fusion model.

## 1 Related Work

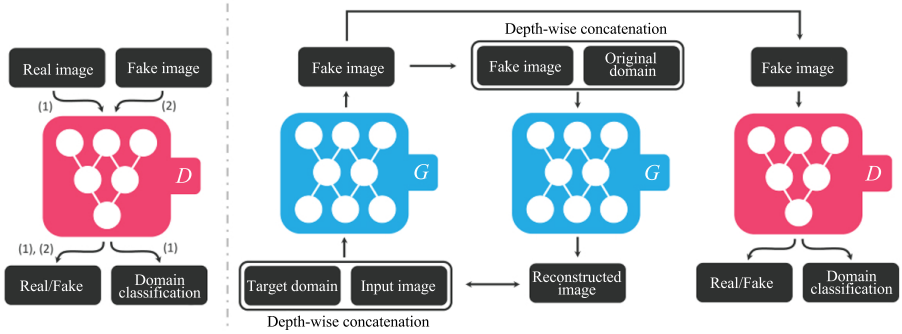
GAN-based image style transfer is a common and popular style transfer method at present due to the powerful generative capacity of GAN and its good effect on image style transfer[20]. Unlike the traditional CNN-based image style transfer, this method can accept a whole class of input images, learn their data distribution and common features, and generate images as similar as possible to the input images on the basis of their features. Differently from the traditional method that learns the style with only one input image, this method can significantly boost the efficiency of image style transfer.

Li and Wang[21] introduced Markov random fields in GAN to train the generative model by adversarial training, which enhanced the realism of the generated images. Then, a series of GAN-based image style transfer models are derived, such as DiscoGAN, DualGAN, Pix2Pix, CycleGAN, and StyleGAN. Among them, DiscoGAN and DualGAN are based on the idea of dual learning of machine translation. In 2017, to seek the cross-domain transfer relationship and solve the problem of model collapse, Kim *et al*.[14] proposed the DiscoGAN model that resorts to the idea of symmetric structure to complete image style transfer. Later, Yi *et al*.[15] proposed the unsupervised image style transfer model DualGAN on the basis of dual learning and L1-norm. In 2017, Isola *et al*.[12] presented the supervised image style transfer model Pix2Pix that adds conditions to the original GAN to control the style of generated images. As a result, they achieved the image style transfer from horses to zebras, apples to oranges, days to nights, grayscale to color images, etc. Pix2Pix improved the generator and discriminator of DCGAN, and it used U-Net to enhance details and PatchGAN to process the high-frequency parts of the images. However, the training data must be paired so that a large number of paired images are needed for the network model training, and paired datasets are difficult to obtain in many cases. In 2018, Zhu *et al*.[13] developed the image style transfer model CycleGAN that uses a cycle consistency loss to enable training without the need of paired data, and it only requires two datasets of input and output. CycleGAN first transfers the image from domain A to domain B and then transfers it from domain B to domain A; through such a cycle, the images before and after the transfer are paired. This process is similar to supervised learning and can significantly advance the transfer effect. In 2019, Karras *et al*.[11] proposed StyleGAN that mainly changes the structure of the generator to achieve the unsupervised generation of highly controllable images. Instead of pursuing more realistic images, this model is dedicated to improving the network's ability to precisely control the generated images. These excellent adversarial training models derived from GAN through continuous improvement have successfully achieved unsupervised image style transfer with constant updates in input data, network structure, loss function, etc.

## 2 Fundamental Theories

### 2.1 StarGAN

The StarGAN network model proposed by Choi *et al*.[22] in 2018 can be used to solve the image style transfer problem among multiple domains, which only requires the training of one generator to complete the multi-domain image style transfer. The network structure of StarGAN is shown in Fig. 1[22], and similarly to the traditional GAN structure, it consists of two modules, namely the generator $G$ and the discriminator $D$. Fig. 1(b) shows the input of the source domain images and target domain labels into the generator to generate fake images, and then the fake images and original domain labels are employed as input to reconstruct and generate original images through the process shown in Fig. 1(c).




(a) Training the discriminator (b) Original-to-target domain (c) Target-to-original domain (d) Fooling the discriminator

**Figure 1** Network structure of StarGAN[22]

The generator does its best to generate images that are indistinguishable from the real images, and the discriminator D discriminates the authenticity of these images and classifies them into the target domain. The adversarial loss function is shown in Eq. (1):

$$L_{\text{adv}}(G,D) = E_x[\log D_{\text{src}}(x)] + E_{x,c}[\log(1 - D_{\text{arc}}(G(x,c)))] \tag{1}$$

where $x$ is the input image and $c$ is the target domain label. The generator generates images according to the target domain label $c$. The reconstructed loss function of the process depicted in Fig. 1(c) is shown in Eq. (2).

$$L_{\text{rec}} = E_{x,c,c'}[\|x - G(G(x,c),c')\|_1] \tag{2}$$

where $c'$ is the source domain label, and the source domain image is reconstructed using the generated fake image and the source domain label. The discriminator $D$ learns to discriminate between real and fake images, and at the same time it is responsible for classifying the input images into the corresponding domains for the multiple classes of input images. The loss of discrimination and the loss of domain classification are shown in Eqs. (3) and (4):

$$L_{\text{cls}}^{f} = E_{x,c}[-\log D_{\text{cls}}(c|G(x,c))] \tag{3}$$

$$L_{\text{cls}}^{r} = E_{x,c'} = [-\log D_{\text{cls}}(c'|x)] \tag{4}$$

The final generator consists of the adversarial loss, the reconstruction loss, and the classification loss, while the discriminator is composed of the adversarial loss and the classification loss, as shown in Eqs. (5) and (6), respectively:

$$L_G = L_{\text{adv}} + \lambda_{\text{cls}} L_{\text{cls}}^{f} + \lambda_{\text{rec}} L_{\text{rec}} \tag{5}$$

$$L_D = -L_{\text{adv}} + \lambda_{\text{cls}} L_{\text{cls}}^{f} \tag{6}$$

where $\lambda_{\text{cls}}$ and $\lambda_{\text{rec}}$ are hyperparameters.

## 2.2 U-Net and edge-promoting adversarial loss function

The most prominent feature of U-Net is the skip connection, which stitches the feature images in the channel dimension to form thicker features. U-Net relies on a superposition method instead of a simple summation operation when fusing shallow features, and it has many feature channels that enable the network to propagate the shallow range information to the high-resolution layer. Despite the fact that the up-sampling can fill in the image information,

it still loses some information, and thus the feature image should be connected with the high-resolution feature image. During up-sampling, the feature image becomes increasingly abstract as the extracted feature images are more efficient and abstract with the increase in convolutional layers. Therefore, it needs to be connected to the high-resolution feature image, which is equivalent to a compromise between the high-resolution feature image and the more abstract feature.

By skip connection, U-Net incorporates much detailed information from the input images, which is conducive to restoring the information loss caused by down-sampling. The up-sampling process fuses the output feature images in down-sampling through skip connection, which, from another point of view, is a fusion of features of different sizes and a kind of multi-scale feature fusion. For example, the last up-sampling has both the output of the first convolutional layer and the large-scale output of up-sampling, and there are four fusion processes involved in the whole network.

The edge-promoting adversarial loss function is proposed to preserve the clear image edges. The task of the discriminator $D$ during the whole training process is to discriminate the authenticity of input images, but it is observed through experiments that it is insufficient to merely train the discriminator $D$ to discriminate between the generated images and the real images. The reason is that clear edges are also an important feature in the image style transfer process, but these edges make up only a small portion of the whole image. Therefore, the discriminator trained with the standard loss function may be confused in the case where there is a correct shading output but no clear reproduction of image edges. Considering this, an edge-promoting adversarial loss function is presented for the discriminator $D$. Specifically, a simple edge removal preprocessing is required for the original input images, and the processed images are then input to the discriminator for discrimination, which can improve the discriminator's ability to discriminate image edges. Once the discriminator detects an anomaly in the edges of the generated images, it gives feedback on this situation to the generator, and then the generator generates higher quality images through iterative optimization. The obtained edge-promoting adversarial loss function is shown in Eq. (7):

$$L_{\text{edge}} = E_{x'}[\log(1 - D(x'))] \tag{7}$$

where $x'$ is the input image of the original dataset after edge removal. The loss function of the final discriminator $D$ is shown in Eq. (8):

$$L_D = -L_{\text{adv}} + \lambda_{\text{cls}} L^r_{\text{cls}} + \lambda_{\text{edge}} L_{\text{edge}} \tag{8}$$

It consists of three parts, namely the adversarial loss function, the classification loss function, and the edge-promoting adversarial loss function, where $\lambda_{\text{cls}}$ and $\lambda_{\text{edge}}$ are the hyperparameters of the domain classification loss and edge loss. The classification loss function $L^r_{\text{cls}}$ is shown in Eq. (9):

$$L^r_{\text{cls}} = E_{x,c'}[-\log D_{\text{cls}}(c'|x)] \tag{9}$$

where $x$ is the input real image and $c'$ is the source domain label; the discriminator $D$ is guided by the source domain label to classify the input images to corresponding domains appropriately.

## 2.3 Class encoder

In the process of image style transfer, the generator extracts the mean and variance of each channel of the feature images, which ultimately affect the styles of the generated images. The generator of UE-StarGAN adopts the Instance Normalization (IN), which is to normalize $H$ and

$W$ to obtain the mean and variance, so as to improve the image style transfer effect and accelerate the convergence of the model. However, UE-StarGAN is limited to style transfer between source domain images and requires massive training samples. Therefore, a class encoder is incorporated to achieve image style transfer with a small sample size. The class encoder serves to map a set of images from $N$ classes to the latent class codes that can be decoded into the mean and variance, and decoding results are input into the generator to affect the image style transfer effect.

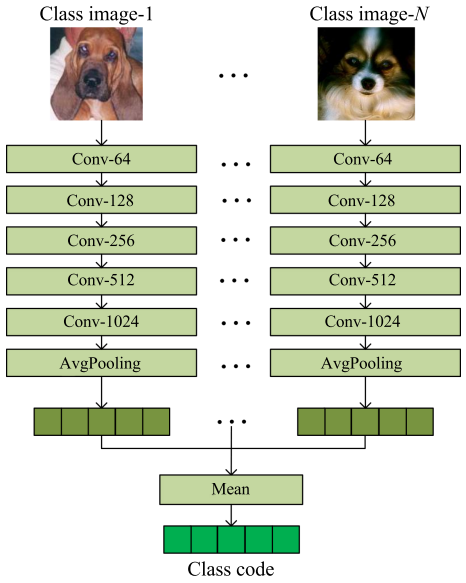



**Figure 2** Network structure of class encoder

The class encoder takes a set of target images as input, and it can use a small number of input target images of different classes to train the model together with the generator. In other words, the class encoder can extract the styles of images, and the generator can extract the texture structure and content of the input source images; finally, the source images are transferred to target class images of different styles. The network structure of the class encoder is shown in Fig. 2, which consists of five convolutional layers and an average pooling layer, and each convolutional layer is followed by a ReLU activation function. The class encoder first maps each image from $N$ classes into $N$ intermediate latent variables, and then the $N$ intermediate latent variables are averaged to generate the final latent class code Zy. To ensure that the generated images have the features of the target images, we propose a feature matching loss function to guarantee the similarity between input and output. The feature matching loss function is given in Eq. (10):

$$L_{\text{FM}}(G) = E_{x,y_1,\cdots,y_k}\left[\left\|F(\bar{x}) - F\left(\sum_k \frac{F(y_k)}{k}\right)\right\|_1\right] \tag{10}$$

where $x$ is the input source image; $\bar{x}$ is the output image; $y_1, \cdots, y_k$ are the input target class images, and $F$ is the feature extractor. $L_1$ regularization is adopted to minimize the feature loss, and thus the similarity between the output image and the target image can be ensured.

# 3 Proposed Models and Algorithms

## 3.1 Image style transfer model based on improved StarGAN

On the basis of StarGAN, a style transfer model based on U-Net and the edge-promoting adversarial loss function, i.e., UE-StarGAN, is constructed to address the problems of background color distortion and blurred image edge details after image style transfer. The GAN model contains a generator and a discriminator: the generator in StarGAN is improved according to the idea of skip connection in the U-Net model, and the edge-promoting adversarial loss function is introduced into the discriminator to improve its ability to discriminate the edge details of the generated images. The structure of the improved generator network is shown in Fig. 3.

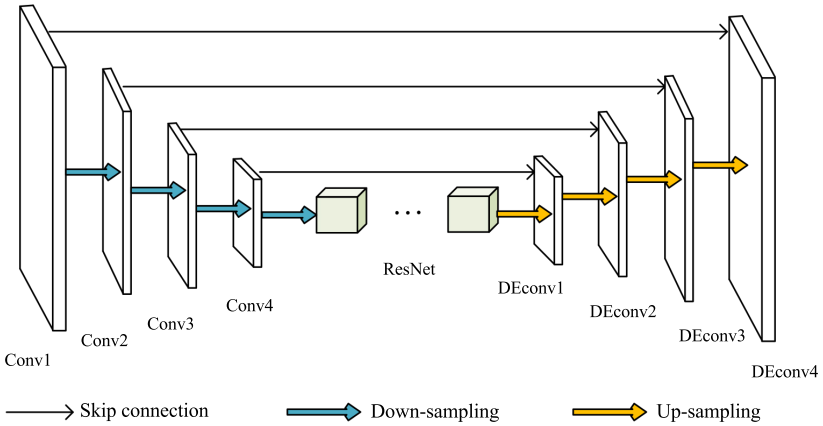



**Figure 3** Structure of the UE-StarGAN Generator Network

As shown in Fig. 3, the generator is comprised of four convolutional layers, six residual modules, and four deconvolutional layers. The fusion of feature images in up-sampling with the deconvolutional feature images through skip connection helps to compensate for the information loss caused by restoring the down-sampling. The adversarial loss function of the generator is shown in Eq. (11):

$$L_{\text{adv}}(G, D) = E_x[\log D_{\text{src}}(x)] + E_{x,c}[\log(1 - D_{\text{src}}(G(x, c)))] \tag{11}$$

where $x$ is the input source domain image and $c$ is the target domain label. The generator can transfer the input images into different styles of images under the guidance of the target label.

## 3.2 Image style transfer model fusing class encoder with a small sample size

Unsupervised image style transfer models achieve great success in improving cross-domain style transfer, but their image style learning from a small number of new-class samples is still limited as the inference based on their prior knowledge is completely beyond their capabilities. These models need to learn the distribution of sample data from tens or hundreds of thousands of new-class images to generate stylized images of the target class, and thus, they do not support image style transfer with a small sample size. To address this problem, this paper builds an image style transfer model fusing class encoder on the basis of UE-StarGAN, namely, CUE-StarGAN. The network model contains two parts, i.e., a generator and a discriminator.

The generator in UE-StarGAN is improved according to the idea of class encoding in class encoder. The structure of the generator network in CUE-StarGAN is shown in Fig. 4, and the

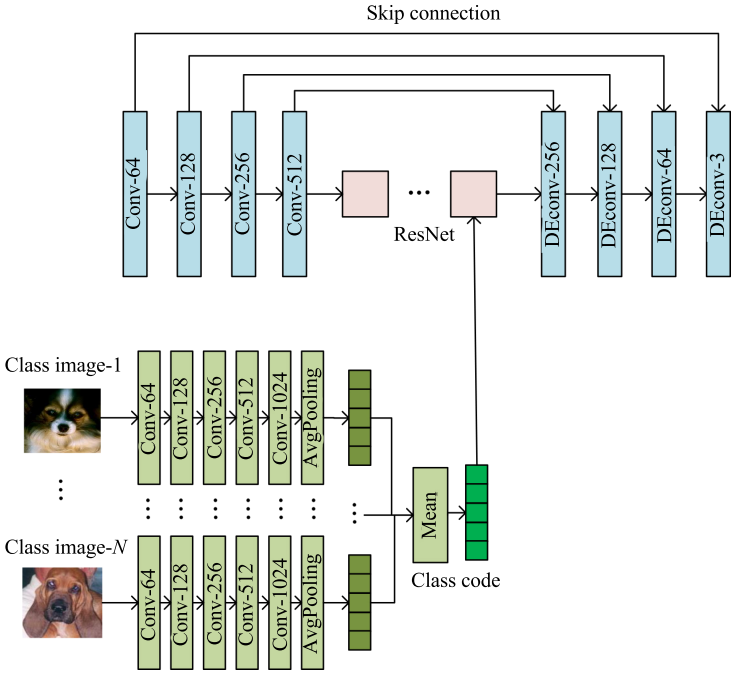



**Figure 4** Structure of CUE-StarGAN generator network

generator can be regarded as an assembly of the UE-StarGAN generator and the class encoder. The source class images and the target class images are the input of the generator network: the main content features of the source class images are extracted by up-sampling, and the latent class codes of the target class images with certain feature styles are extracted by the class encoder. Finally, the latent class codes are taken to the decoder through the residual modules to control the image style. The residual module consists of adaptive instance normalization (AdaIN) and residual blocks, where AdaIN normalizes the sample data in each channel to zero mean and unit variance and then obtains the global appearance information by affine transformation to control the style of the generated images.

The loss function of the generator contains the adversarial loss function, the reconstruction loss function, and the feature matching loss function. The adversarial loss function and the reconstruction loss function are shown in Eq. (12) and Eq. (13), respectively:

$$L_{\text{adv}} = E_x[\log D(x)] + E_{x,y_1,\cdots,y_k}[\log(1 - D(\bar{x}))] \tag{12}$$

$$L_R = E_{x,c}[\|x - G(\bar{x}, c')\|_1] \tag{13}$$

where $x$ represents the input source image; $\bar{x}$ stands for the output image; $y_1, \cdots, y_k$ represent the input target class images; and $c'$ is the source domain label. The adversarial loss function of the final generator is shown in Eq. (14).

$$L_G = L_{\text{adv}} + \lambda_R L_R + \lambda_{\text{FM}} L_{\text{FM}} \tag{14}$$

where $\lambda_R$ and $\lambda_{\text{FM}}$ are the hyperparameters of the reconstruction loss function and the feature matching loss function, respectively.

# 4 Experimental Results and Analysis

## 4.1 Datasets

The datasets used in the experiments and tests in this paper are CelebA[22] and Fer2013, and the Animal Faces and North American Birds datasets in ImageNet. CelebA is the abbreviation of CelebFaces Attribute, which is the authoritative and complete attribute dataset for celebrity faces in the field of face recognition and facial expression research; it contains more than 200,000 images of faces with different attributes from a total of more than 10,000 celebrities. Each image in the dataset has attribute annotations and contains 40 types of face attribute annotations, the face bbox annotation boxes, and five point coordinates of facial features. The human facial expression dataset Fer2013 is a grayscale image dataset mainly used for the study of human facial expression changes, which includes seven types of facial expressions corresponding to the number tags of 0–6 and involves 35,886 images for the seven facial expressions. Among them, 28,708 images are for testing, and the separate number of public validation images and private validation images is 3,589, all with a size of $48\times48$. The Animal Faces and North American Birds datasets are datasets of animal faces and North American birds contained in the ImageNet dataset, respectively, and the ImageNet dataset is a collection of more than 14 million images that covers over 20,000 species, which is created under the leadership of Professor Li Feifei at Stanford University.

## 4.2 Evaluation metrics

The commonly used evaluation metrics[24] for image style transfer are PSNR[25], SSIM[26], IS, and FID. Among them, PSNR and SSIM directly quantify the quality of images, while IS and FID are usually used to evaluate the quality and diversity of images. In this paper, the above four metrics are taken as the quantitative evaluation metrics for experiments.

PSNR refers to the Peak Signal-to-Noise Ratio, which can evaluate and measure the distortion and noise of two images. A larger PSNR value indicates less distortion and the high realism of generated images. SSIM refers to the Structural Similarity Index, which measures the similarity between two images in terms of structure, contrast, and brightness; a larger SSIM means the generated images are more realistic in terms of structure, brightness, and contrast.

IS represents the Inception Score, which evaluates the quality of the generated images in terms of clarity and diversity using only the relevant information of the generated data; a greater IS represents higher diversity and better quality of the generated images. FID refers to the Fréchet Inception Distance, which measures the performance of a network in terms of image diversity and quality and determines the quality of the generated images by calculating the distance between the real data and the generated data at the feature layer. A smaller FID value indicates that the generated data is closer to the real data, and the quality of the generated images is higher.

## 4.3 Experimental results

To verify the effectiveness of the proposed model, experiments are conducted on AIOS, a deep learning platform based on the Linux operating system and that supports RTX 2080 Ti, 4-core CPU, 16 GB GPU, 24 GB RAM, and PyTorch framework for GPU-accelerated computation. In the training process, the batch processing size is 16, and the number of iterations is 200,000; the generator adopts IN; the learning rate of both the generator and discriminator is 0.0001, and the Adam optimizer is used for training.

Firstly, CelebA and Fer2013 datasets are applied to train the UE-StarGAN model proposed in Section 3.1. Then, 2,000 images are randomly selected from CelebA and Fer2013 datasets separately for testing, and comparisons are made with other image style transfer model algorithms

including Pix2Pix, CycleGAN, and StarGAN on different datasets. SSIM and PSNR are selected as evaluation metrics, and the mean of the test results obtained on the two datasets are given in Tables 1 and 2.

**Table 1** Comparison of SSIM and PSNR results on CelebA

| CelebA | SSIM | PSNR (dB) |
|---|---|---|
| Pix2Pix | 0.767 | 21.463 |
| CycleGAN | 0.749 | 20.686 |
| StarGAN | 0.788 | 22.752 |
| **UE-StarGAN** | **0.881** | **25.653** |

**Table 2** Comparison of SSIM and PSNR results on Fer2013

| Fer2013 | SSIM | PSNR (dB) |
|---|---|---|
| Pix2Pix | 0.834 | 25.085 |
| CycleGAN | 0.859 | 24.107 |
| StarGAN | 0.866 | 25.882 |
| **UE-StarGAN** | **0.879** | **26.262** |

Tables 1 and 2 give the SSIM and PSNR values on the CelebA dataset with glasses, blond hair, fringes, gender, and age as well as on the Fer2013 dataset with five different facial expressions including normal, angry, happy, sad, and surprised. Both SSIM and PSNR values of the UE-StarGAN model are higher than those of the other four models, which demonstrates that the UE-StarGAN model outperforms other image style transfer models on the whole. The visualized comparison results on Fer2013 are shown in Fig. 5.


**Figure 5** Comparison results on Fer2013

In Fig. 5, (a) represents the Pix2Pix model; (b) represents the CycleGAN model; (c) represents the StarGAN model, and (d) represents the proposed UE-StarGAN model. By a horizontal comparison, the proposed UE-StarGAN shown in (d) can generate images with higher resolution, and the facial expressions upon transfer are more natural and effective while retaining the basic facial content information. The Pix2Pix model is not as good as the other three models, especially for mouths and eyes that are blurred upon transfer, and the processing of the details is rough. By a vertical comparison of the images in the columns of angry, sad, and surprised, the images in (a), (b), and (c) are all blurred at the eyes, noses, and mouths, while the

images generated by the proposed method are clearer and more delicate in the detailed features and edges.

To verify the performance of the CUE-StarGAN model proposed in Section 3.2, we train and test the model on the Animal Faces and North American Birds sub-datasets of th ImageNet datasets. Then, it is compared with other image style transfer models of CycleGAN, UNIT, StarGAN, and UE-StarGAN on different datasets, and IS and FID are selected as evaluation metrics. IS is a metric for measuring the individual and overall features of the generative model, and a larger IS denotes that the generated images are more realistic, and the transfer effect is better; FID is a principled and comprehensive metric, and a lower FID indicates that the quality of the generated images is higher, and the image diversity is richer. The test results obtained on the two datasets are given in Tables 3 and 4.

**Table 3** Comparison results on Animal Faces

| Animal faces | IS | FID |
|---|---|---|
| CycleGAN[27] | 7.43 | 197.13 |
| UNIT[27] | 12.14 | 197.13 |
| StarGAN[27] | 6.21 | 198.07 |
| **UE-StarGAN** | **8.96** | **186.78** |
| **CUE-StarGAN** | **13.75** | **165.49** |

**Table 4** Comparison results on North American Birds

| North American birds | IS | FID |
|---|---|---|
| CycleGAN[27] | 25.28 | 215.30 |
| UNIT[27] | 28.28 | 203.83 |
| StarGAN[27] | 18.94 | 260.04 |
| **UE-StarGAN** | **23.76** | **230.5** |
| **CUE-StarGAN** | **37.43** | **197.86** |

Tables 3 and 4 demonstrate the comparison with the classical CycleGAN, UNIT, and StarGAN models on Animal Faces and North American Birds datasets. Considering the results, it can be concluded that the UE-StarGAN algorithm proposed in Section 3.1 outperforms the original StarGAN model. In addition, the IS and FID values of the CUE-StarGAN model proposed in Section 3.2 are better than those of the other models, where the IS value is significantly enhanced, and the FID value is lower than all that of other algorithms. In particular, the CUE-StarGAN fusing class encoder shows a remarkably enhanced effect, which reveals that the algorithm proposed can generate more realistic images with a small sample size and richer image diversity upon the style transfer and verifies the effectiveness of the CUE-StarGAN model.

Next, the visualization of the model on the Animal Faces and North American Birds sub-datasets of ImageNet is presented, as shown in Figs. 6 and 7.

Columns (a) and (b) in Figs. 6 and 7 represent the input class images; column (c) provides the input source images, and column (d) represents the images upon transfer by CUE-StarGAN. It can be seen from Fig. 6 that CUE-StarGAN can transfer a Shiba Inu into a white pet dog while basically retaining the overall content information of the Shiba Inu and resembling the class images in style. Fig. 7 shows that the overall content information of an owl basically remains, and the style resembles the class images; CUE-StarGAN transfers the owl into images with a North American bird style. The above visualization results further verify the effectiveness of the proposed model that can achieve the transfer of image styles with a small number of class samples.

(a) (b) (c) (d)

**Figure 6** Visualization of the model on animal faces of ImageNet

(a) (b) (c) (d)

**Figure 7** Visualization of the model on North American birds of ImageNet

## 5 Conclusion

In this paper, the classical image style transfer model StarGAN is improved on the basis of GAN from two aspects, namely the detailed feature extraction and the small sample size. Firstly, the skip connection in the U-Net module is introduced and combined with the StarGAN module for feature extraction and generation of images, and the edge-promoting adversarial loss

function is added to the discriminator to enhance its discrimination capability when dealing with edge features of images. Thus, the image style transfer model UE-StarGAN based on the improved StarGAN is presented. Secondly, the class encoder is investigated on the basis of UE-StarGAN. The style features of the input class images are learned through the class encoder, and the image style transfer is completed using a few class images. The image style transfer model CUE-StarGAN with a small sample size is then proposed. The experimental results reveal that the model has the advantage in the case of a small sample size and can extract finer features and better achieve the style transfer from source images to class images. The images subjected to the image style transfer have shown improvement in both qualitative and quantitative analyses, which verifies the effectiveness of the proposed model.

## References

[1] Ligeza A. Artificial intelligence: A modern approach. Applied Mechanics & Materials, 2009, 263(2): 2829–2833.

[2] Pollack ME. Artificial intelligence—A modern approach (a review). AI Magazine, 1995, 16: 73–74.

[3] Zhou FY, Jin LP, Dong J. Review of convolutional neural network. Chinese Journal of Computers, 2017, 40(6): 1229–1251 (in Chinese with English abstract).

[4] Lu HT, Zhang QC. Applications of deep convolutional neural network in computer vision. Journal of Data Acquisition and Processing, 2016, 31(1): 1–17 (in Chinese with English abstract).

[5] Li S. Research and development of natural language processing. Journal of Yanshan University, 2013, 37(5): 377–384 (in Chinese with English abstract).

[6] Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. In: Ghahramani Z, ed. Proc. of the Advances in Neural Information Processing Systems. MIT Press, 2014. 2672–2680.

[7] Zhang D, Shao J, Hu G, *et al*. Sharp and real image super-resolution using generative adversarial network. Proc. of the Int'l Conf. on Neural Information Processing. Cham: Springer International Publishing, 2017. 217–226.

[8] Ouyang N, Liang T, Lin LP. Self-attention network based image super-resolution. Journal of Computer Applications, 2019, 39(8): 2391–2395 (in Chinese with English abstract).

[9] Gao Y, Liu Z, Qin PL, *et al*. Medical image super-resolution algorithm based on deep residual generative adversarial network. Journal of Computer Applications, 2018, 38(9): 2689–2695 (in Chinese with English abstract).

[10] Reed S, Akata Z, Yan X, *et al*. Generative adversarial text to image synthesis. JMLR.org. 2016.

[11] Karras T, Laine S, Aila T. A Style-based generator architecture for generative adversarial networks. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2019. 4401–4410.

[12] Isola P, Zhu JY, Zhou TH, *et al*. Image-to-image translation with conditional adversarial networks. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2017. 5967–5976.

[13] Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv: 1703.10593, 2017.

[14] Kim T, Cha M, Kim H, *et al*. Learning to discover cross-domain relations with generative adversarial networks. Proc. of the 34th Int'l Conf. on Machine Learning (ICML). IMLS, 2017. 2941–2949.

[15] Yi Z, Zhang H, Gong PTM. DualGAN: Unsupervised dual learning for image-to-image translation. arXiv: 1704.02510, 2017.

[16] Ratliff LJ, Burden SA, Sastry SS. Characterization and computation of local Nash equilibria in continuous games. Proc. of the 51st Annu. Allerton Conf. on Communication, Control, and Computing (Allerton). 2013. 917–924.

[17] Chen Y, Lai YK, Liu YJ. CartoonGAN: Generative adversarial networks for photo cartoonization. Proc. of the IEEE/CVF Conf. on Computer Vision & Pattern Recognition. IEEE, 2018. 9465–9474.

[18] Qian R, Tan RT, Yang W, *et al*. Attentive generative adversarial network for raindrop removal from

a single image. Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018. 2482–2491.

[19] Yao NM, Guo QP, Qiao FC, *et al*. Robust facial expression recognition with generative adversarial networks. Acta Automatica Sinica, 2018, 44(5): 865–877 (in Chinese with English abstract).

[20] Chang JY. Image style transferring based on generative adversarial network [MS. Thesis]. Xuzhou: China University of Mining and Technology, 2021 (in Chinese with English abstract).

[21] Li C, Wand M. Precomputed real-time texture synthesis with Markovian generative adversarial networks. Proc. of the European Conf. on Computer Vision. Cham: Springer, 2016. 702–716.

[22] Choi Y, Choi M, Kim M, *et al*. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. arXiv: 1711.09020, 2017.

[23] Yao Z, Zhang BY, Wang ZY. IntersectGAN: Learning domain intersection for generating images with multiple attributes. Proc. of the ACM Int'l Conf. ACM, 2019. 1842–1850.

[24] Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004, 13(4): 600–612.

[25] Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 2008, 44(13): 800–801.

[26] Zhu XS, Yao SR, Sun B, *et al*. Image quality assessment: Combining the characteristics of HVS and structural similarity index. Journal of Harbin Institute of Technology, 2018, 50(5): 121–128 (in Chinese with English abstract).

[27] Liu MY, Huang X, Mallya A, *et al*. Few-shot unsupervised image-to-image translation. Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). IEEE, 2019. 10550-10559.

**Xinzheng Xu**, Ph.D., professor, CCF senior member. His research interests include machine learning, data mining, and pattern recognition.

**Shifei Ding**, Ph.D., professor, CCF senior member. His research interests include machine learning, data mining, and pattern recognition.

**Jianying Chang**, master candidate. Her research interests include deep learning and computer vision.