Received: 9 June 2022 | Revised: 8 December 2022 | Accepted: 9 December 2022

DOI: 10.1002/cav.2135

RESEARCH ARTICLE

WILEY

FAEC-GAN: An unsupervised face-to-anime translation based on edge enhancement and coordinate attention

Hong Lin | **Chenchen Xu** | **Chun Liu**

Wuhan University of Technology, Wuhan, China

Correspondence
Chun Liu, Wuhan University of Technology, Wuhan, China.
Email: 1950380336@qq.com

Abstract

Animation is a widely loved artistic form with high abstraction and powerful expression. The task of image translation from face to anime involves complex geometric and texture transformations, and requires the generated images with clear lines. The existing unsupervised image translation frameworks are often ineffective for this task. According to the characteristics of animation image, we propose an animation translation method based on edge enhancement and coordinate attention, which is called FAEC-GAN. We design a novel edge discrimination network to identify the edge features of images, so that the generated anime images can present clear and coherent lines. And the coordinate attention module is introduced in the encoder to adapt the model to the geometric changes in translation, so as to produce more realistic animation images. In addition, our method combines the focal frequency loss and pixel loss, which can pay attention to both the frequency domain information and pixel information of the generated image to improve the visual effect of the image. The experimental results demonstrate that FAEC-GAN is superior to the state-of-the-art methods in the task of face-to-animation image translation.

KEYWORDS

computer vision, deep learning, generative adversarial networks

1 | INTRODUCTION

Animation is a widely loved art form, which is highly abstract and expressive. Anime avatars are often used as personal avatars on social platforms to reflect personal preferences and characteristics. However, it is very difficult and time-consuming to draw corresponding anime faces according to the facial features of different individuals. Therefore, it is a meaningful work to generate the corresponding anime images according to the face photos.

It is hard for the previous image style transfer method[1–3] to generate anime faces from real photos, because the style transfer method can only transfer the style information (texture, color style, etc.) of images. However, there are complex geometric changes in the task of generating animated faces from real photos. But the image translation methods[4–6] can accomplish this task well, since the image is generated from the feature maps, instead of modifying it based on the structure of the source image. The task of image translation is to transfer the input image from the source domain to the target domain, and make the translated image have the characteristics of the target domain. As shown in Figure 1, our method uses the image translation structure based on GAN to complete the translation task from face to anime face.

At present, many unsupervised methods can generate anime images from real photos. CartoonGAN[7] first achieved excellent results in the task of photo cartoonization. But it only focuses on the transformation of picture texture details

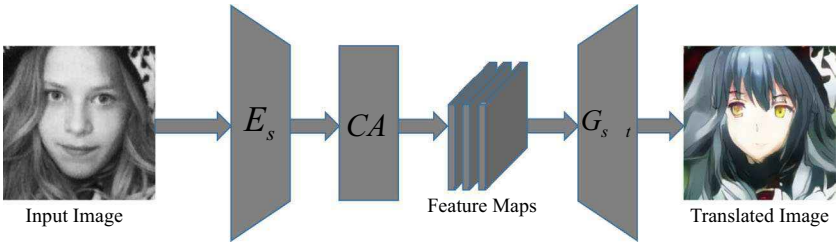

FIGURE 1 Our model translates real photos into anime faces. E_s is the encoder of the source image domain, and CA is the coordinate attention module. $G_{s\to t}$ is a generator, which can generate the target image from the feature maps.

and styles. In the translation task from face to anime, CartoonGAN cannot adjust the translation strategy according to the contour features of the anime face, which leads to the significant difference between the generated anime images and the real anime works.

U-GAT-IT[8] uses class activation map (CAM)[9] to adapt the geometric changes between domains in image translation. By using channel attention mechanism, U-GAT-IT can focus on the differences between the source domain and the target domain, which can well solve the problems brought by large shape changes in image domains. As a result, U-GAT-IT has achieved better results than the previous methods in the task of generating anime faces from photos. Chen[10] proposes a more compact structure for the image translation task, which showed great results in the task of generating anime faces. AniGan[11] proposes a novel generator that can transfer both color and texture styles, which can generate corresponding anime image based on the structure of the source image.

Although showing superior effects, there are still three limitations in the above image translation methods. First, the current image translation methods are not designed exclusively for the characteristics of anime faces. The generated anime images of them are often messy and incoherent. Second, these methods do not focus on the key parts of the feature maps in both space and channel dimensions, which makes it difficult to generate realistic images. Third, these methods often only focus on the pixel-wise difference between real images and generated images, but ignore the frequency-domain difference between them, which leads to the decline of the visual effect of generated images.[12,13]

In order to overcome these limitations, we propose an animation translation method based on edge enhancement[14] and coordinate attention, called FAEC-GAN, to complete the translation task from real photos to anime faces. First of all, we find that clear and coherent lines play a decisive role in the quality of animation pictures, which is often ignored by previous methods. Therefore, we propose an edge discrimination network, which consists of an edge detection module and an edge discriminator. The edge information of the images are extracted by the edge detection module and sent to the edge discriminator for discrimination. In this way, the image generated by the generator is forced to contain clear and coherent lines.

Second, there are great differences between the real faces and the anime faces, and the existing models often do not pay enough attention to the inter-domain differences, resulting in poor quality of generated images. For this reason, we first introduce the coordinate attention mechanism[15] into the image translation task. The structure of generator and discriminator is adjusted to focus on the key parts of feature maps accurately. In addition, inspired by FFL,[16] we find that frequency information plays a key role in image generation, while the previous methods only pay attention to the pixel differences of images, and ignore the frequency differences between images. Therefore, we design a novel loss function, which combines focal frequency loss and spatial loss to measure the difference between images. In this way, the generators can obtain both the spatial information and frequency information of images, and learn the distribution of real data.

Our main contributions are as follows:

1. We propose a novel edge discrimination network to discriminate the edge features of images, enabling the generated anime images with clear and coherent lines
2. The coordinate attention module that combines the spatial attention and the channel attention is used in the encoder. This enables our method to focus on the inter-domain differences and improve the translation quality
3. We combine focal frequency loss and pixel loss to measure the difference between the generated images and the original images, which makes the generated image more realistic
4. Experiments show that each module is effective. In addition, compared with the most advanced methods, our method has achieved better results in the task of face to anime

2 | RELATED WORK

2.1 | GAN-based image-to-image translation

Image translation model can be used to complete many image processing tasks, such as image super-resolution, image inpainting, style transfer and so on.[17] After generative adversarial nets (GAN)[18] was proposed, the image translation models based on GAN[19–21] are widely used in computer vision because of its excellent effect. Pix2pix[4] uses paired data sets for training, which can convert one type of image to another, and achieves great results in various translation tasks. Pix2pixHD[22] is based on Pix2pix, which can synthesizes high-resolution images.

The problem of obtaining paired data is solved by unsupervised image translation methods. CycleGAN[23] is the first to complete the unsupervised image translation task by using cycle consistency constraint and paired generator and discriminator, which reduces the dependence on paired datasets. Dual-GAN[24] has a similar structure of CycleGAN. When updating the parameters of discriminator, it will not only use the currently generated images, but also use the previously generated images for cooperative training, to solve the problem of mode collapse. Considering the unbalanced information in different image domains, Yi and Liu[25] propose an asymmetric cycle structure to complete the task of generating portraits from face images. For the task of converting real photos into cartoon images, CartoonGAN[7] uses semantic loss and edge loss to ensure that the image content will not be changed and the edges will be clear.

However, when these methods are used for the tasks with great geometric changes, the image quality is obviously degraded. In order to solve this problem, Wu et al.[6] propose a landmark assisted GAN for cartoon face generation, which uses extra manual annotation information to assist the model to locate the facial structure. AttentionGAN[26] uses attention mask to distinguish the foreground and background of the image, so as to focus on the change of the foreground. U-GAT-IT[8] designs a learnable normalization function AdaLIN and focuses on the difference between image domains by the channel attention mechanism to obtain images with higher quality. NICE-GAN[10] proposes a novel structure to complete the task of image translation. It reuses the encoder of the discriminator making the model compact and efficient. ACL-GAN[27] proposes an adversarial consistency loss, which let the generated images retain the important features of the source images, rather than all the information of them. In this way, artifacts can be avoided in the generated images. SPatchGAN[28] is an asymmetric cycle structure, and it discriminates key statistical features on multi-scale, so that the model can adapt to the shape change between image domains well.

2.2 | Attention mechanisms

Attention mechanism[29] is a method to simulate human attention, which can make the model set different weights to different parts of the input. Therefore, attention mechanism can help the model focus on the parts that are more important to the results. Attention mechanism has been widely used in deep learning tasks such as natural language processing and image recognition since it was proposed. Jie et al.[30] proposed a channel attention mechanism called squeeze-and-excitation networks, which obtains the global spatial feature of each channel by squeezing, and learns the interdependencies between channels by excitation. Compared with previous methods, it gets significantly improvement on many tasks and datasets. Similarly, by using the channel attention mechanism, class activation mapping[9] shows the areas that neural networks focus on in the task of classification. CBAM[31] uses the sequence structure of channel attention and spatial attention to obtain the attention map, which effectively helps the model to pay attention to the key information in the feature maps. The coordinate attention method proposed by Hou et al.[15] further improves the performance of attention mechanism in various classic networks. This method pools the input feature map along the vertical and horizontal directions to obtain two independent attention maps, which are used to jointly save the spatial information and channel information of the input feature map.

3 | OUR FAEC-GAN

The goal of our method is to get a mapping from face images to anime faces by using unpaired datasets. Considering that animation images contain coherent lines and clear edges, we propose an unsupervised face-to-anime translation model based on edge enhancement and coordinate attention, which is called FAEC-GAN. In the following discussion, let X_s represent the face domain and X_t represent the anime face domain.

3.1 | Architecture

As shown in Figure 2, our model consists of two generators $G_{s\to t}$, $G_{t\to s}$ and four discriminators D^s_{adv}, D^t_{adv}, D^s_{edg}, D^t_{edg}. $G_{s\to t}$ can convert the input real faces into anime images. Correspondingly, $G_{t\to s}$ can convert the input anime into face. D^t_{adv} discriminates whether the anime face image is real or generated, and D^s_{adv} is similar to it. D^s_{edg} and D^t_{edg} discriminate the input image edge information to make the generators generate images with realistic edge details. We reuse the encoders E_s, E_t of multi-scale discriminators D^s_{adv}, D^t_{adv}, and for the first time introduce the idea of coordinate attention[15] in image translation. We add the coordinate attention module (CA) after the encoder, which helps the model pay attention to the key information in the feature map by combining the spatial features and channel features of the images. Besides, we design a novel loss function for cycle-consistency loss and reconstruction loss, which combines focal frequency loss[16] and pixel loss to measure the difference between images accurately. In this way, the generators can obtain both the spatial information and frequency information of images, and learn the distribution of real data. In this section, we only describe the training process from real face domain X_s to anime face domain X_t, because the training process from X_t to X_s is symmetrical.

3.1.1 | Discriminators based on coordinate attention

The previous image translation methods cannot well adapt to the huge geometric difference between the face images and the anime images. In order to solve this problem, we introduce the CA module into the discriminators D^s_{adv} and D^t_{adv}. As shown in Figure 3, the coordinate attention module pools the input feature map in the vertical and horizontal directions to obtain two independent attention maps, which are used to jointly store the spatial and channel information of the input feature maps. The original feature maps and attention maps are combined and sent to the classifier C_s to discriminate the input image. By combining spatial attention and channel attention, the coordinate attention module urges the model to focus on the key parts of the feature maps that affect the results significantly and improves the accuracy of discrimination.

Since the structure of our model is symmetrical, we will only discuss D^s_{adv} here. Let $x \in \{X_s, G_{t\to s}(X_t)\}$ represent the source images or images translated from the target domain. After the input image x is encoded by encoder E_s, the feature map $E_s(x)$ is obtained. The CA module pools the input feature map $E_s(x)$ in the vertical and horizontal directions, to obtain attention maps with spatial and channel information. Then the attention map is combined with the input feature map to get a new feature map CA ($E_s(x)$). Finally, CA ($E_s(x)$) is sent to a multi-scale discriminator to discriminate the input image.

We use multi-scale discriminators to judge the authenticity of the generated image, and reuse the encoder in discriminators as the encoders of the generators. Our multi-scale discriminator discriminates the different scales of input

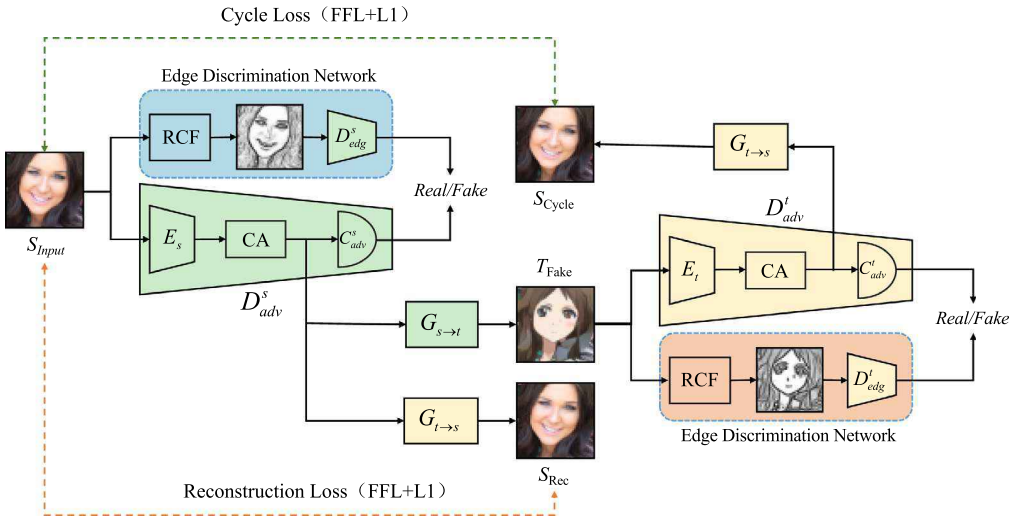

FIGURE 2 Overall structure of FAEC-GAN. Here, we only show the translation process from the real face domain S to the anime face domain T. S_{Input} is the input face image, T_{Fake} is the generated anime image, S_{Rec} is the self-reconstruction image of S_{Input}, and S_{Cycle} is the reconstruction image of S_{Input}

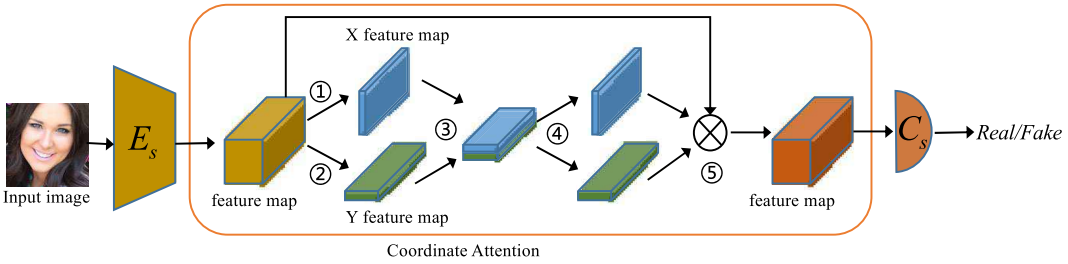

FIGURE 3 Structure of discriminator D^s_{adv}. After the processing of encoder and coordinate attention module, the input image is sent to the classifier C_s to obtain the result. Here, ① is pooling in the X direction, ② is pooling in the Y direction, ③ is concatenation and convolution, ④ is batch regularization and separation, and ⑤ is multiplying the original feature maps with the obtained attention maps

images by down-sampling, so as to improve the accuracy of judging the image. In addition, since the generators and the discriminators contain similar encoders, sharing encoders can make the model more compact, which reduces the number of parameters and improves the training effect. Concretely, traditional training of the encoder is conducted by back-propagating the gradients from the generator, which is indirect. By reusing the encoder, it can be trained by the loss function directly, which is more compact and effective.

3.1.2 | Edge discrimination network

Coherent lines and clear edges are the distinctive features of anime works. However, compared with hand-drawn images, the anime images generated by the current image translation method have messy lines and blurred edges. To solve this problem, we designed an edge discrimination network to discriminate the edge information of the input images, so as to make the generator to generate images with clear edges.

In order to compare the edge information of real and generated images, we extract edge information from real animation works and translate anime images by using trained RCF edge detection model.[32] As shown in Figure 4, the edges of real animation are coherent and neat, while the edges of the generated anime faces are often broken and blurred. It shows the necessity of enhancing the edge information of the generated images.

As shown in Figure 5, the edge discrimination network consists of an edge discriminator and a RCF edge detection network. The edge discriminator is used to judge whether the input edge image comes from the real image or the generated image. RCF is an edge detection network, which can detect the edge features of input images and output corresponding edge images. RCF network is divided into five stages, which extract edge information of different scales. Take stage 3 for example, the feature map output by stage 2 is pooled and sent to stage 3. And the feature information is extracted by convolution layer and then sent to stage 4. At the same time, the deconvolution layer is used to output the edge image of stage 3. In this way, the information in each receptive field is used to extract richer edge information.

Since the edges only account for a tiny part of the whole image, we extract the edge information of the image by the trained RCF edge detection module, and then send it to an independent edge discriminator. As a result, it is avoided to input the complete image to the discriminator, which will cause the discriminator to be disturbed by a large amount of useless information. In this way, the edge discriminator can directly discriminate the edge information of the image, forcing the generator to focus on the edge of the input and generate anime images with smooth lines.

3.1.3 | Generators

Our generators and discriminators share a pair of encoder E_s and E_t, and generate the target image by the input hidden vectors. Let $x \in \{X_s, G_{t\to s}(X_t)\}$ represent the source images or images translated from the target domain. As shown in Figure 2, after the input image is encoded by encoder E_s, the feature map $E_s(x)$ is obtained. Then the feature map CA $(E_s(x))$ is obtained by CA module. Finally, the generator $G_{s\to t}$ generates the target image from CA $(E_s(x))$. In addition, we use AdaLIN[8] to dynamically select the regularization method for the generators.

FIGURE 4 Comparison of edge extraction. (a) Is a real anime work and its edge extraction. (b) Is the generated anime face and its edge extraction

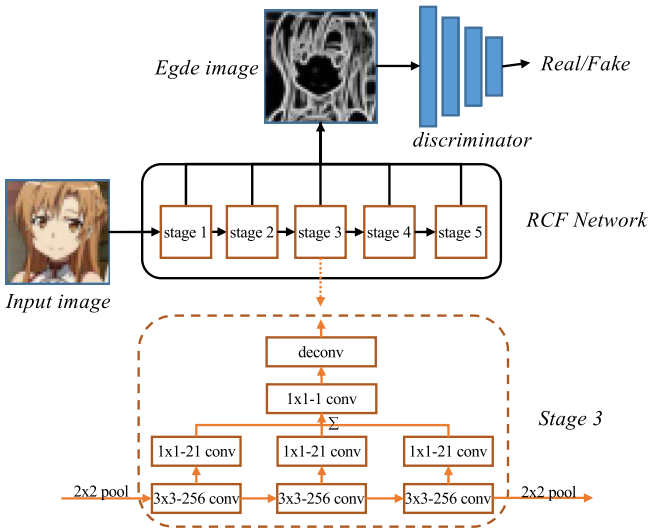

FIGURE 5 Structure of edge discrimination network. The edge discrimination network consists of RCF edge detection module and discriminator. The RCF network consists of five stages, which acquire edge features of different scales. We expand stage 3 to show its details

3.2 | Loss function

The training of our model is guided by five loss functions. We explain them in detail as follows:

Focal frequency loss: Focal frequency loss (FFL) is designed to make the model focus on frequency components that are hard to synthesize. The existing methods[8,10] usually complete images translation tasks by using pixel losses. They compare two images pixel by pixel, and calculate the difference between them. But each coordinate value on the frequency spectrum is determined by the pixels of the whole image. Therefore, it is difficult to reduce the frequency differences between two images by comparing them pixel by pixel, especially some frequency components that are difficult to synthesize (i.e., hard frequencies). But FFL can extract the frequency information of images directly, and

calculate the differences between them in frequency domain. So FFL is complementary to the existing spatial losses. By using both FFL and spatial losses, images can be evaluated in multiple dimensions, thus making the image more realistic.

Focal frequency loss can calculate the frequency differences between images and make the model focus on hard frequencies adaptively. To obtain the frequency information of the images, we perform the 2D discrete Fourier transform as shown in Equation (1), to convert the image to its frequency representation. Here the image size is $M \times N$; (x, y) denotes the coordinate of an image pixel in the spatial domain; $f(x, y)$ is the pixel value; (u, v) represents the coordinate of a spatial frequency on the frequency spectrum; $f(v, u)$ is the complex frequency value; i is the imaginary unit.

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}. \tag{1}$$

To focus the model on the hard frequencies, we use a spectrum weight matrix to adjust the weight of each frequency. The spectrum weight matrix is dynamically determined by the current loss of each frequency during training. The matrix element $w(v, u)$, that is, the weight for the spatial frequency at (v, u) is shown in Equation (2). Here α is the scaling factor for flexibility. And we normalize the matrix values into the range $[0, 1]$, where the weight 1 corresponds to the currently most lost frequency.

$$w(u,v) = \left|F_r(u,v) - F_f(u,v)\right|^{\alpha}. \tag{2}$$

Finally, the full form of the FFL is obtained by performing a Hadamard product of the spectrum weight matrix and the frequency distance matrix:

$$\mathrm{FFL} = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w(u,v) \cdot \left|F_r(u,v) - F_f(u,v)\right|^2. \tag{3}$$

Adversarial loss: By using the adversarial loss, the generator learns the mapping from the source domain to the target domain. Therefore, the generated images are as close as possible to the target images. It is formulated as:

$$L_{adv}^{s\to t} = \mathrm{E}_{x\sim X_t}\left[\left(D_{adv}^{t}(x)\right)^2\right] + \mathrm{E}_{x\sim X_s}\left[\left(1 - D_{adv}^{t}\left(G_{s\to t}\left(CA\left(E_s(x)\right)\right)\right)\right)^2\right]. \tag{4}$$

Here, CA is the attention module. E_s is the encoder that converts the source image into the hidden vector. D_t is the discriminator, which composed of the classifier C_t and the encoder E_t, to judge whether the input image is real. In order to prevent the generator and discriminator from interacting with each other, we use decoupled training strategy. When training the discriminator D_t, we fix $G_{s\to t}$ and E_s and update the parameters of C_t and E_t. When training the generator, we fix E_s and E_t, and only update the parameters of $G_{s\to t}$.

Edge loss: In order to match the edge discrimination network proposed in Section 3.1.2, we designed an edge loss to learn the edge details of real images to obtain better anime images. Similar to the adversarial loss, edge loss is adversarial. The input image is processed by the RCF module to get the edge image, and then the edge discriminator judges its authenticity. The edge loss is formulated as:

$$L_{\mathrm{edg}}^{s\to t} = \mathrm{E}_{x\sim X_t}\left[\left(D_{\mathrm{edg}}^{t}(\mathrm{RCF}(x))\right)^2\right] + \mathrm{E}_{x\sim X_s}\left[\left(1 - D_{\mathrm{edg}}^{t}\left(\mathrm{RCF}\left(G_{s\to t}\left(\mathrm{CA}\left(E_s(x)\right)\right)\right)\right)\right)^2\right] \tag{5}$$

Here, RCF is the edge detection module, which is used to obtain the edge image of the input image. D_{edg}^{t} is an edge discriminator to judge whether the edge image is real. Other settings are similar to the adversarial loss.

Cycle-consistency loss: Cycle-consistency loss[23] is used to ensure the consistency of image content. In addition, L1 loss and FFL loss are combined to measure the similarity of images. By comparing their frequency representations, the loss function makes the generator pay attention to the frequency information of the generated images, which is complementary to existing spatial losses. The cycle-consistency loss is formulated as:

$$L_{\mathrm{cycle}}^{s\to t} = \mathrm{E}_{x\sim X_s}\left[\lambda_L \left|x - G_{t\to s}\left(G_{s\to t}(x)\right)\right|_1 + \lambda_{\mathrm{FFL}} \left|x - G_{t\to s}\left(G_{s\to t}(x)\right)\right|_{\mathrm{FFL}}\right], \tag{6}$$

where $|\cdot|_1$ is L1 loss, and $|\cdot|_{\text{FFL}}$ is FFL loss. λ_L and λ_{FFL} are weights of L1 loss and FFL loss. We set $\lambda_L = 1$ and $\lambda_{\text{FFL}} = 10$ to balance L1 loss and FFL loss.

Reconstruction loss: Reconstruction loss is used to retain the identity information. In addition, L1 loss and FFL loss are combined to measure the similarity of images. The reconstruction loss is formulated as:

$$L_{\text{rec}}^{s\to t} = \mathrm{E}_{x\sim X_s}\left[\lambda_L \left|x - G_{t\to s}(x)\right|_1 + \lambda_{\text{FFL}} \left|x - G_{t\to s}(x)\right|_{\text{FFL}}\right]. \tag{7}$$

Here, settings are similar to the cycle-consistency loss.

Full objective: Finally, we jointly train all components to optimize the final objective:

$$\min_{G_{s\to t}, G_{t\to s}} \max_{D_{\text{adv}}^s, D_{\text{adv}}^t} \lambda_1 L_{\text{adv}} + \min_{G_{s\to t}, G_{t\to s}} \max_{D_{\text{edg}}^s, D_{\text{edg}}^t} \lambda_2 L_{\text{edg}} + \lambda_3 L_{\text{cycle}} + \lambda_4 L_{\text{rec}}. \tag{8}$$

Here, $L_{\text{adv}} = L_{\text{adv}}^{s\to t} + L_{\text{adv}}^{t\to s}$, and other losses ($L_{\text{edg}}$, L_{cycle}, and L_{rec}) are defined in the similar way. λ_1, λ_2, λ_3, and λ_4 are the weights of adversarial loss, edge loss, cycle-consistency loss and reconstruction loss. We set $\lambda_1 = 1$, $\lambda_2 = 0.5$, $\lambda_3 = 5$, and $\lambda_4 = 5$ to balance losses. Note that the weight of edge loss is set to 0.5, because when it is too large, the generator will pay too much attention to the edge information of the generated images. As a result, the model will ignore the identity information of the images, which leads to a great difference between the color of the input and output images.

4 | EXPERIMENTS

4.1 | Baselines

FAEC-GAN is compared with state-of-the-art methods, including CycleGAN,[23] U-GAT-IT,[8] NICE-GAN,[10] ACL-GAN,[27] and SPatchGAN.[28] All the baseline methods are from the public codes and set to the parameters provided by the papers.

CycleGAN puts forward the cycle consistent loss. That is, the image obtained by inverse mapping from the generated image, should be as close as possible to the input image. In this way, CycleGAN is the first to get wonderful results in unsupervised image translation task.

U-GAT-IT introduces an attention module to lead the model get the key parts of feature maps. And a new normalization function AdaLIN is designed to select instance normalization (IN) and layer normalization (LN) adaptively. Due to the limitation of GPU, we use its light version for comparison.

NICE-GAN proposes a compact structure to translate image in unsupervised ways. Considering that the generators and discriminators often use similar encoders, this method reuses the encoders in discriminators and get great performance. Due to the limitation of GPU, we use its light version for comparison.

ACL-GAN propose an adversarial consistency loss, which let the generated images retain the important features of the source images, rather than all the information of them. In this way, artifacts can be avoided in the generated images.

SPatchGAN proposes an asymmetric structure, which makes the model better adapt to the shape change between image domains. In addition, its discriminator focuses on the statistical features of the image, rather than the local features of the image, which makes the network more stable.

In addition, we compared U-GAT-IT-light and NICE-GAN-light with our FAEC-GAN, and the total number of parameters and FLOPs of network modules are shown in Table 1. It shows that FAEC-GAN has similar parameters and FLOPs, even if additional edge discrimination networks are used.

4.2 | Datasets

The goal of our method is to get high-quality anime images from real face photos. For this purpose, we use the following datasets for training and testing.

Selfie2anime: It was first used in U-GAT-IT.[8] It contains 3500 selfie images and 3500 anime faces, in which 3400 selfie images are used for training and 100 selfie images are used for testing. The division of anime image is similar. The images of selfie2anime are unpaired. In addition, all the selfie images and anime images are from female characters, and resized to 256×256 by using the super-resolution algorithm.

TABLE 1 Total number of parameters and FLOPs of network modules. Since U-GAT-IT-light and NICE-GAN-light do not contain the edge discrimination networks, the corresponding parameters and FLOPs are all 0

Module / Method	Number of params (FLOPs)			
	Generators	Discriminators	Edge networks	Total
U-GAT-IT-light	21.2 M (105.0G)	112.8 M (15.8G)	0.0 M (0.0G)	134 M (120.8G)
NICE-GAN-light	11.5 M (48.2G)	93.7 M (12.0G)	0.0 M (0.0G)	105.2 M (60.2G)
FAEC-GAN	11.4 M (48.3G)	93.6 M (11.7G)	20.3 M (31.7G)	125.4 M (91.6G)

FIGURE 6 Example real faces and anime faces of our ce2anime dataset

Ce2anime: In order to evaluate the effect of the proposed method, we established an additional face2anime dataset which is called ce2anime, some are shown in Figure 6. It contains 10,000 real faces and 10,000 anime faces, in which 9500 face images are used for training and 500 images are used for testing. The division of anime image is similar. 10,000 real face images are selected from CelebA[33] randomly and resized to 256×256. Ten thousand anime character images are selected from Danbooru2018[34] randomly. And we obtain 10,000 anime faces from those anime characters images, which are resized to 256×256.

4.3 | Evaluation metrics

We use Frechet inception distance (FID)[35] and the Kernel inception distance (KID)[36] to evaluate the quality of generated images. FID is used to calculate the similarity between two groups of images, and it is often used to evaluate the quality of images generated by generative adversarial network. The lower the FID, the more realistic the images are.

Similar to FID, KID is used to measure the difference between two groups of images by calculating their statistical features. Unlike the FID, KID is an unbiased metric, which is more consistent with human perception. The lower the KID, the more realistic the images are.

4.4 | Setup

Our method is implemented in PyTorch. And we train baseline methods on NVIDIA RTX 2080Ti GPU to obtain the experimental results. All experiments are trained by using Adam optimizer, and we set $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is set to 0.0001. We use ReLU as the activation function in the generator and leaky-ReLU as the activation function in the discriminator. Since the ce2anime contains more diversiform images than the self2anime, it needs more iterations to converge. Considering the convergence of each model on different datasets, all models are trained for 200 k iterations on self2anime and trained for 300 k iterations on ce2anime. The training takes about 19.8 h per 100 k iterations. The batch size of all experiments is set to 1. For data augmentation, we flipped the images with a probability of 0.5 and resized them to 286×286 and randomly cropped to 256×256.

4.5 | Ablation study

We conduct ablation experiments to validate the effectiveness of individual components in our method: (1) edge discrimination network, (2) coordinate attention module, (3) the loss function that combines focal frequency loss and pixel loss.

4.5.1 | Edge discriminant network analysis

In order to validate the effectiveness of the edge discriminant network, we delete the edge discriminant network in FAEC-GAN and compare it with the complete FAEC-GAN model. As shown in Figure 7, we show the (a) source images, (b) results with EDG network and (d) results without EDG network. Besides, in order to highlight the role of the edge discrimination network in our method, additional edge images (c) and (e) are obtained from (b) and (d) by the edge

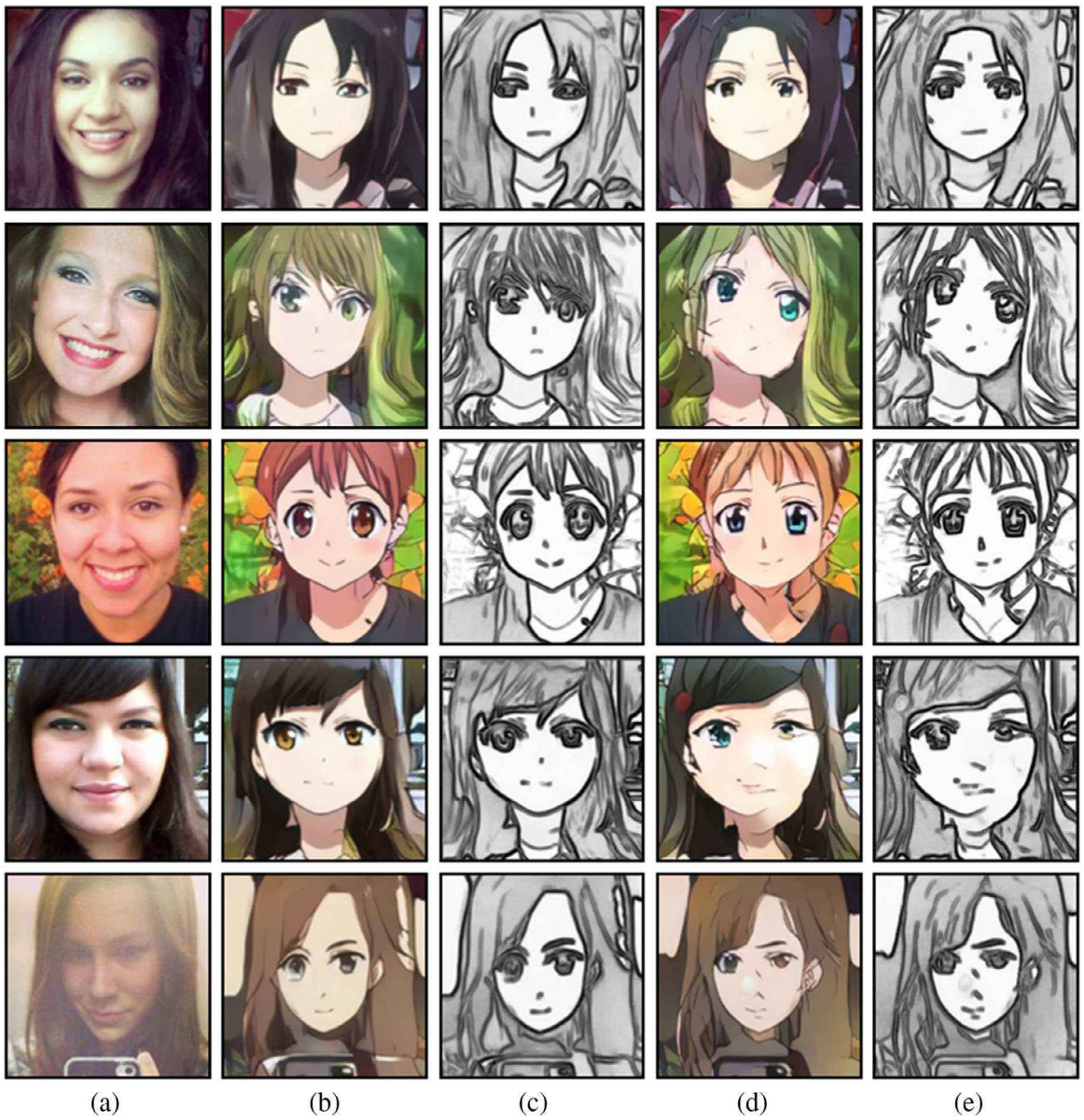

FIGURE 7 Comparison of the results with or without EDG network: (a) source images, (b) our results, (c) edge images of (b), (d) results without EDG network, (e) edge images of (d)

detection model. As shown in (b) and (d), in row 1, the edges of the anime face in (b) are coherent and clear, while some messy spots appear on the faces in (d). In the rows 2 and 4, the edges of the anime face with EDG network are clearly, while the lines that without EDG are broken in the face. In addition, we find that the images generated by the method with edge discrimination network have more symmetrical eyes (as shown in Figure 7, rows 2, 3, and 4). Because the edge discrimination network can extract the line of the image directly, which better shows the structure of the face, and the coordinate attention makes the model focus on the key parts of the image. By using both of them, our method can generate realistic images. To sum up, the edge discrimination network can generate anime faces with clear and coherent lines, and significantly improve the quality of them.

4.5.2 | Coordinate attention analysis

We conduct four groups of experiments to validate the effectiveness of the coordinate attention module. First, we remove the CA module in FAEC-GAN and compare it with the complete FAEC-GAN model. In addition, we replace CA with CAM and CBAM to verify the effectiveness of different attention methods in our FAEC-GAN. As shown in Figure 8c, the model without CA does not perform well, and the content of the source image is not preserved well (e.g., hair and pupil color of the character are obviously deviated from the source images in row 1). As shown in Figure 8d,e, compared with the results without attention module, the quality of the generated images with CAM and CBAM models has been greatly improved, but there are still some defects (e.g., in Figure 8d, a part of the hair in row 3 turns red and some artifacts appear on the face in row 2 of Figure 8e). In contrast, the images generated by FAEC-GAN contain fewer artifacts, clearer edges and the best visual effect. For faces with different poses(rows 3 and 4), the method with coordinate attention can capture

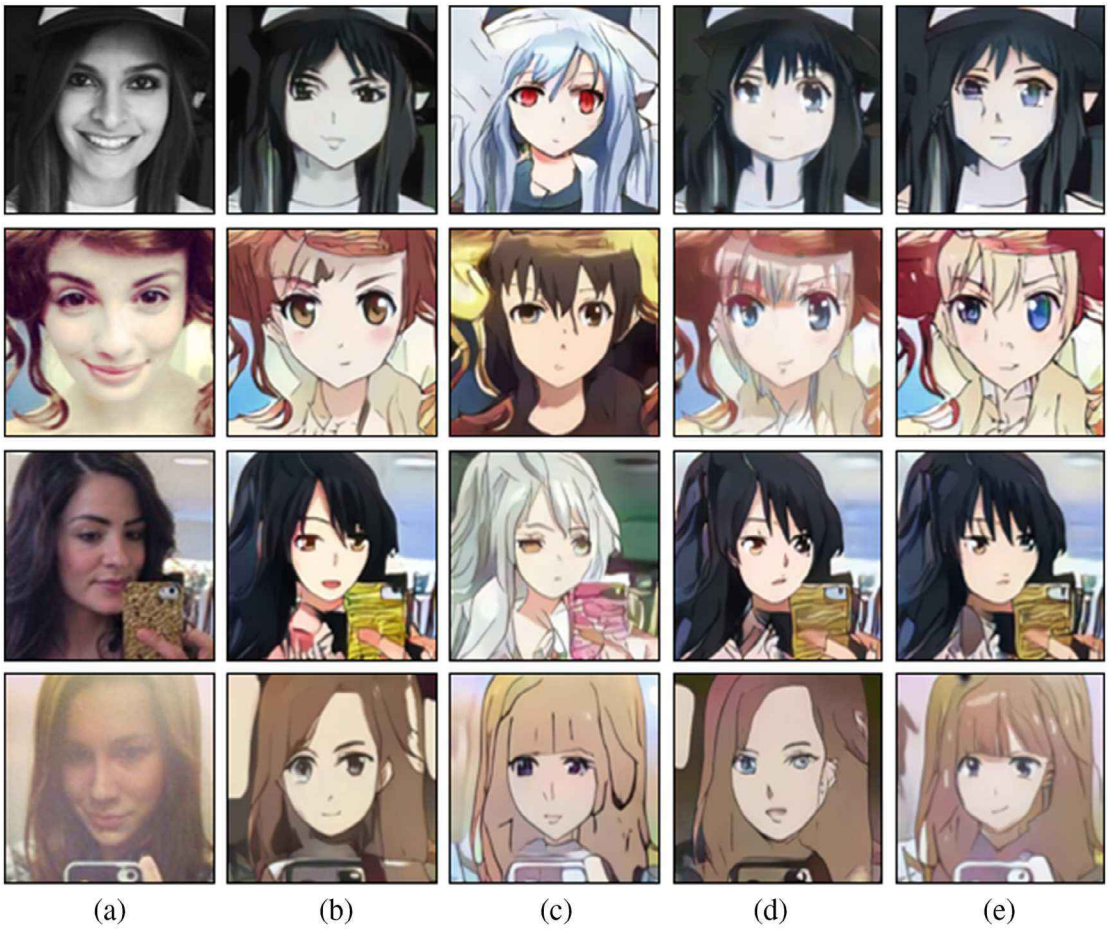

FIGURE 8 Comparison of the results with different attention module: (a) source images, (b) our results, (c) results without CA, (d) results with CAM, (e) results with CBAM

the facial structure better and retain the information of the input images. Overall, the coordinate attention module can help our model to focus on the key areas accurately and improve the translation effect.

4.5.3 | Loss function analysis

We design a novel loss function for cycle-consistency loss and reconstruction loss, which combines focal frequency loss and pixel loss to measure the difference between images accurately. In order to verify the effectiveness of this loss function, we replace it with focal frequency loss and pixel loss in FAEC-GAN, and compare them with the FAEC-GAN. As shown in Figure 9b, the model combining L1 loss and FFL loss can well retain the information (including structure and color) of the source images. But the methods using only FFL or L1 loss are not as satisfactory (e.g., in Figure 9c row 3, the eyes are asymmetrical and the edges are not clear, and in Figure 9d row 2 the face is distorted). To sum up, the results show that L1 loss and FFL loss are complementary and can obtain the pixel and frequency information of images well.

In addition, to further validate the key factors in our method, as shown in Table 2, we quantitatively evaluated the above seven experiments using the FID and KID. It can be seen that our FAEC-GAN method achieves the best results, while the other control models all show varying degrees of performance degradation.

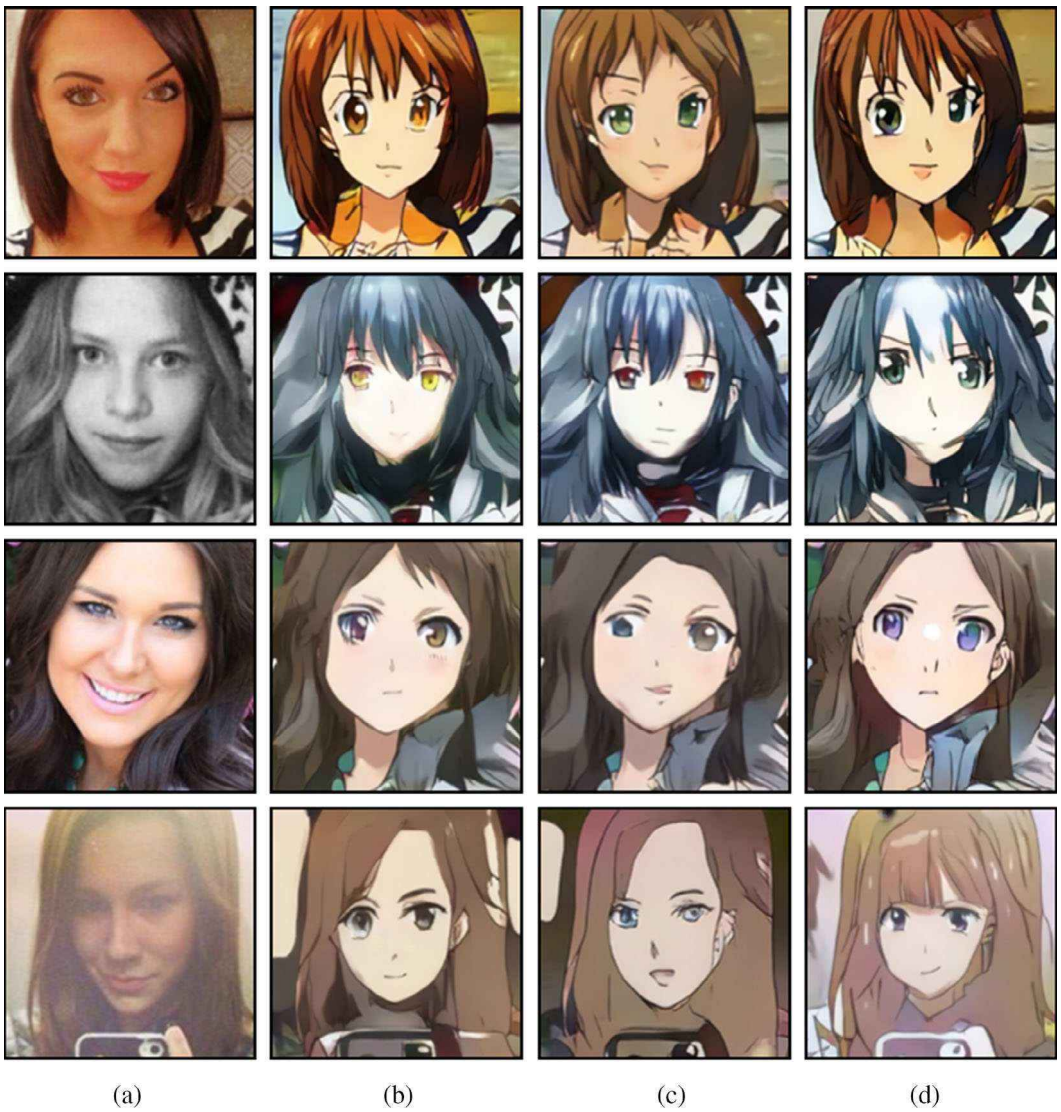

FIGURE 9 Comparison of the results using different loss functions in cycle-consistency and reconstruction: (a) source images, (b) our results, (c) results only using FFL loss, (d) results only using L1 loss

TABLE 2 Ablation study. EDG is the edge discrimination network we proposed. L1 is pixel loss and FFL is focal frequency loss. CAM is an attention module named class activation map. CBAM is convolutional block attention module and CA is the coordinate attention module. "w/" means that the component is used in the experiment, "w/o" means that the component is not used. The lower the FID and KID, the better the results

Metric / Model	FID	KID × 100
FAEC-GAN	92.92	2.76
FAEC-GAN w/o EDG	98.61	3.47
FAEC-GAN w/L1	94.14	3.23
FAEC-GAN w/FFL	97.45	3.42
FAEC-GAN w/o CA	103.09	4.31
FAEC-GAN w/CAM	98.72	3.46
FAEC-GAN w/CBAM	101.48	3.96

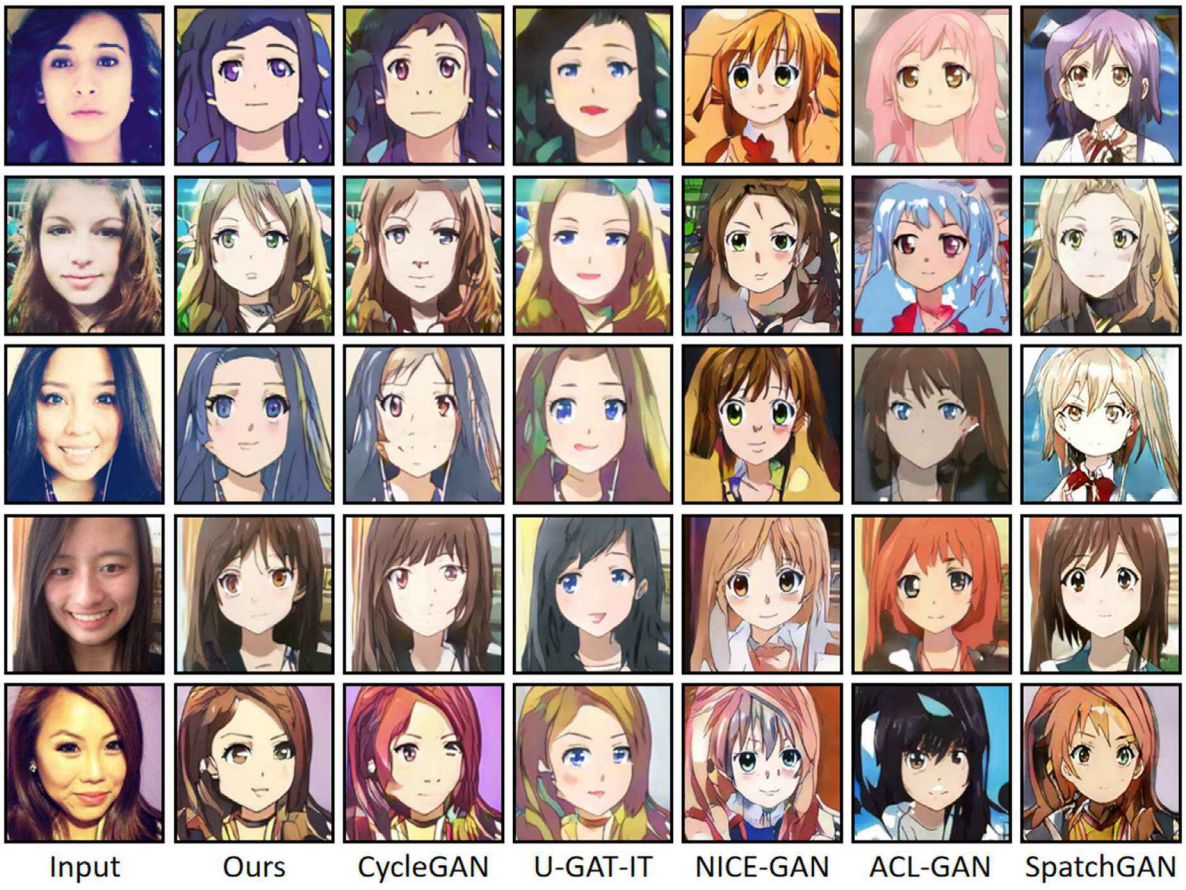

FIGURE 10 Generated images of FAEC-GAN and the baselines on self2anime

4.6 | Comparison with baselines

In this section, we compare our FAEC-GAN with the baselines (Section 4.1) on the self2anime and ce2anime (Section 4.2).

First, we compare the performance of each model on self2anime. As shown in Figure 10, CycleGAN can well imitate the texture and color of the target images without using attention module. But it results in more noise and artifacts.

The images generated by U-GAT-IT are more realistic, but there are still some defects in details (e.g., the missing lines between the lip and the eyes, and the inconsistent size of the eyes in row 4). NICE-GAN does not focus on the key parts of the input images (e.g., some hair is translated into the background in the first row). ACL-GAN and SPatchGAN can generate images with high quality, but it cannot keep the content of the original images well, which may be due to the asymmetric structure used in this method. For example, in row 1 and row 3, compared with the source images, the color of the images generated by ACL-GAN and SPatchGAN changed obviously. And extra clothing appears in row 3 of the SPatchGAN. Overall, our FAEC-GAN can retain the content of the source images well and generate realistic anime faces.

To further validate the FAEC-GAN, we compare our FAEC-GAN with the baselines on the ce2anime. The generated images of FAEC-GAN and the baselines on ce2anime are shown in Figure 11. For female faces, the images generated by our method contain sharper lines than other methods and retain the information of the source image better (e.g., in row 4, the blonde hair color and facial orientation of the input image are retained). For faces with different poses (rows 4 and 5), our method is able to generate anime faces with the corresponding poses, preserving the information of the source images well. For faces of different races (rows 3 and 5), our method is able to generate anime faces that most closely resembled the source images. Overall, our FAEC-GAN can retain the content of the source images and generate realistic anime faces when dealing with real faces of different races and poses.

In addition, in order to further validate the effectiveness of our model, as shown in Table 3, we evaluate above experiments quantitatively by the FID and KID. On the self2anime, compared with ACL-GAN, which performs best among baselines, our method reduces the FID by 1.95 and the KID by 0.37. On the ce2anime, compared with SpatchGAN, our method reduces the FID by 3.03 and the KID by 0.61. It can be seen that our FAEC-GAN has achieved the best results on different datasets. It shows that our method can learn the distribution of the different datasets and generate realistic anime

FIGURE 11 Generated images of FAEC-GAN and the baselines on ce2anime

TABLE 3 Quantitative metrics of FAEC-GAN and the baselines. Lower is better

Model \ Dataset	Self2anime		Ce2anime	
	FID	KID × 100	FID	KID × 100
FAEC-GAN	92.92	2.91	60.28	2.71
Cycle-GAN	114.54	3.80	70.90	3.67
U-GAT-IT	105.78	3.93	68.28	3.21
NICE-GAN	112.62	5.41	68.04	3.61
ACL-GAN	94.87	3.28	63.69	2.81
SpatchGAN	98.78	3.71	63.31	3.32

faces. Moreover, our method has achieved the lowest scores on both metrics, which fully demonstrates our FAEC-GAN reasonably performs well regardless of what measure we have used.

5 | CONCLUSION

In this article, we propose FAEC-GAN, an animation translation method based on edge enhancement and coordinate attention, which is designed for the image translation task from face to anime. We propose an edge discrimination network to discriminate the edge features of images, which can make the generated anime faces have clear and coherent lines. Pixel loss and focal frequency loss are combined to measure the difference of images and guide the training of models. In addition, the coordinate attention module is introduced to make the model to focus on the key parts of feature maps and learn to adapt to the shape change during translation. The effect of each of the key factors we proposed are confirmed in ablation study. Compared with the state-of-the-art methods, our method has lower FID and KID, and achieves better translation performance.

Nevertheless, there is still much we can do to improve our work. For example, there are some artifacts in generated images. This may be due to the fact that real images contains more complex background textures and colors than anime images, while the model tends to focus on anime faces. As a result, the generated image has poor detail at the boundary between the face and the background. In addition, due to the differences between the real faces and anime faces, the cycle-consistency loss may embed some useless information in the generated images, which makes the anime images unrealistic. To address these issues, we will explore semantic loss and asymmetric cycle mapping (by using relaxed cycle consistency loss) in the future.

ORCID

Chun Liu https://orcid.org/0000-0003-0762-1080

REFERENCES

1. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016.
2. Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017;1501-10.
3. Liu S, Lin T, He D, Li F, Wang M, Li X, et al. Adaattn: revisit attention mechanism in arbitrary neural style transfer. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2021.
4. Isola P, Zhu J Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. IEEE; 2016.
5. Song G, Luo L, Liu J, Ma WC, Lai C, Zheng C, et al. AgileGAN: stylizing portraits by inversion-consistent transfer learning. ACM Trans Graph. 2021;40(4):1-13.
6. Wu R, Gu X, Tao X, Shen X, Tai YW, Jia J. Landmark assisted CycleGAN for cartoon face generation; 2019.
7. Chen Y, Lai YK, Liu YJ. CartoonGAN: generative adversarial networks for photo Cartoonization. Proceedings of the IEEE/CVF Conference on Computer Vision & Pattern Recognition. IEEE; 2018.
8. Kim J, Kim M, Kang H, Lee K. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation; 2019.
9. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016.

10. Chen R, Huang W, Huang B, Sun F, Fang B. Reusing discriminators for encoding towards unsupervised image-to-image translation. IEEE; 2020.
11. Li B, Zhu Y, Wang Y, Lin CW, Ghanem B, Shen L. Anigan: style-guided generative adversarial networks for unsupervised anime face generation. IEEE Trans Multimed. 2021;24:4077-91.
12. Rahaman N, Baratin A, Arpit D, Draxler F, Lin M, Hamprecht FA, et al. On the spectral bias of neural networks; 2018.
13. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision – ECCV 2020. Cham: Springer; 2020.
14. Guo Y, Liu Y, Oerlemans A, Lao S, Wu S, Lew MS. Deep learning for visual understanding: a review. Neurocomputing. 2016;187:27-48.
15. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design; 2021.
16. Jiang L, Dai B, Wu W, Loy CC. Focal frequency loss for image reconstruction and synthesis. Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2021.
17. Yu X, Chen Y, Liu S, Li G. Multi-mapping image-to-image translation via learning disentanglement. arXiv preprint arXiv:1909.07877; 2019.
18. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets [C]. Neural Information Processing Systems. MIT Press; 2014.
19. Choi Y, Uh Y, Yoo J, Ha JW. Stargan v2: diverse image synthesis for multiple domains. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2020.
20. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2019.
21. Nizan O, Tal A. Breaking the cycle-colleagues are all you need. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2020.
22. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2018, p. 8798-8807.
23. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision. IEEE; 2017, p. 2223-2232.
24. Yi Z, Zhang H, Tan P, Gong M. Dualgan: unsupervised dual learning for image-to-image translation. Proceedings of the IEEE International Conference on Computer Vision. IEEE; 2017, p. 2849-2857.
25. Yi R, Liu YJ, Lai YK, Rosin PL. Unpaired portrait drawing generation via asymmetric cycle mapping. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2020, p. 8217-8225.
26. Tang H, Liu H, Xu D, Torr PHS, Sebe N. Attentiongan: unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Trans Neural Netw Learn Syst. 2021;1-16.
27. Zhao Y, Wu R, Dong H. Unpaired image-to-image translation using adversarial consistency loss. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. European Conference on Computer Vision. Cham: Springer; 2020. p. 800-15.
28. Shao X, Zhang W. SPatchGAN: a statistical feature based discriminator for unsupervised image-to-image translation. arXiv preprint arXiv:2103.16219; 2021.
29. Mnih V, Heess N, Graves A. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. ACM; 2014, p. 2204-2212.
30. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2019, p. 7132-7141.
31. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer; 2018.
32. Liu Y, Cheng MM, Hu X, Bian JW, Zhang L, Bai X, et al. Richer convolutional features for edge detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2017, p. 3000-3009.
33. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. Proceedings of the IEEE International Conference on Computer Vision. IEEE; 2015, p. 3730-3738.
34. https://www.gwern.net/Danbooru2018
35. Heusel M, Ramsauer H, Unterthiner T, Nessler B. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. Adv Neural Inf Process Syst. 2017;30:6629-40.
36. Bińkowski M, Sutherland D J, Arbel M, Gretton A Demystifying mmd gans. arXiv preprint arXiv:1801.01401; 2018.

AUTHOR BIOGRAPHIES

Hong Lin received the B.S degree in Computer Science and Technology from Wuhan University of Technology, in 1986 and the M.S. degree in Computer Application Technology from Wuhan University of Technology, in 1997. Her research interests include data mining and computer vision.

Chenchen Xu received the B.S degree in Computer Science and Technology from Wuhan University of Technology, in 2020. He is currently working toward the M.S degree in Computer Science and Technology with the school of Wuhan University of Technology. His research interests include deep learning and computer vision.

Chun Liu received a Doctor's degree from the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, in 2015. She has been a visiting scholar to the Research Computing Centre in the University of Queensland from 2019 to 2020. Her research interests include data mining, parallel computing and computer vision.

How to cite this article: Lin H, Xu C, Liu C. FAEC-GAN: An unsupervised face-to-anime translation based on edge enhancement and coordinate attention. Comput Anim Virtual Worlds. 2023;e2135. https://doi.org/10.1002/cav.2135