Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

# Facial expression transformation for anime-style image based on decoder control and attention mask

Xinhao Rao [a], Weidong Min [b,c,d,*], Ziyang Deng [b], Mengxue Liu [b]

[a] School of Software, Nanchang University, Nanchang 330047, China
[b] School of Mathematics and Computer Science, Nanchang University, Nanchang 330031, China
[c] Institute of Metaverse, Nanchang University, Nanchang 330031, China
[d] Jiangxi Provincial Key Laboratory of Virtual Reality, Nanchang 330031, China

ARTICLE INFO

*Keywords:*
Decoder control
Expression transformation
Image-to-image translation
Muti-domain translation
Anime-style image

ABSTRACT

Human facial expression transformation has been extensively studied using Generative Adversarial Networks (GANs) recently. GANs have also shown successful attempts in transforming anime-style images. However, current methods for anime pictures fail to refine the expression control efficiently, leading to control effects weaker than expected. Moreover, it remains challenging to maintain the original anime face identity information while transforming. To address these issues, we propose an expression transformation method for anime-style images. In order to enhance the control effect of discrete emoticon tags, a mapping network is proposed to map them to high-dimensional control information, which is then injected into the network multiple times during transformation. Additionally, for better maintaining the anime face identity information while transforming, an integrated attention mask mechanism is introduced to enable the network's expression control to focus on the expression-related features, while avoiding affecting the unrelated features. Finally, we conduct a large number of experiments to verify the validity of the proposed method, and both quantitative and qualitative evaluations are carried out. The results demonstrate the superiority of our proposed method compared to existing methods based on multi-domain image-to-image translation.

## 1. Introduction

The rapid development of computer vision and graphics has shown its great performance on image processing field such as traffic signs and vehicles recognition [1,2]. Inspired by these studies about the images of real world, researches in the field of anime have been driven. Growing speed of the anime industry is very fast these days. The anime industry is not only important, but also one of the pillar industries in some countries. The total annual market is more than tens of billions or even hundreds of billions of dollars. Considering the heavy workload of designing anime characters for each anime series, creating and designing them with automatic methods can significantly reduce the cost. However, mastering painting skills requires great efforts. At the same time, the existing image editing software and algorithms cannot produce satisfactory anime results. The automatic generation of anime characters provides an opportunity to create customized characters without professional skills, which is very helpful for artists and can save a lot of time.

Image-to-image translation aims to transfer an image from one domain to another while the characteristics of given image remaining unchanged. Recently, image-to-image translation has made some success in agriculture and medical field [3,4]. The generative adversarial network (GANs) [5] proposed by Goodfellow et al. is an unsupervised learning technology, which has achieved amazing success in the task of image generation. GANs have been used by some recent studies to provide some successful attempts for high-quality anime-style image generation. These studies have shown good results in anime content generation, converting real-world pictures into anime-style pictures and automatic coloring of anime sketches. However, research on facial expression transformation of anime is not satisfactory. In particular, the existing methods are unable to effectively control details such as eyebrows and lips during the transformation process. Meanwhile, the identification information, such as the shape of the nose, the color of the skin and the color of the eyes, is frequently lost during the

* Corresponding author.
*E-mail addresses:* raoxinhao@email.ncu.edu.cn (X. Rao), minweidong@ncu.edu.cn (W. Min), 359100220003@email.ncu.edu.cn (Z. Deng), mx17522@email.ncu.edu.cn (M. Liu).

https://doi.org/10.1016/j.image.2025.117343
Received 28 April 2024; Received in revised form 17 December 2024; Accepted 1 May 2025
Available online 6 May 2025

transformation process.

Fig. 1 shows some examples of the failure to control expressions and the failure to retain the original information during the transformation process. It can be seen that the transformed images have difference in hair color and skin color, and the effect of changing the expressions is unstable. The complexity and color texture diversity of anime faces are far higher than those of real faces, also there are small differences between classes but great differences within classes. For example, for real human beings, we can easily tell the difference in appearance between two people from the size and position of their features and shapes, but for anime face pictures, the difference in features is hard to notice. Anime faces are highly simplified and abstract, and different anime works are influenced by the author's different drawing styles, brush strokes, and other factors. These intricately-distributed data determine that the difficulty of anime face expression transformation task is far greater than that of real face. Anime face cannot use continuous action units with intensity to represent facial muscle movement, so it cannot be trained with the dataset marked by facial expression coding system (FACS) like the real facial expression transformation method. The existing methods are based on multi-domain image-to-image translation model and training on the dataset marked by discrete expression tags, which still has some problems. On the one hand, the discrete expression label has a single value. After simply copying and expanding it into control information, the redundant value can hardly refine the expression control; moreover, the multi-domain image-to-image translation model used has weak control over expression features. On the other hand, the existing studies use the method of cycle consistency loss to maintain the identity information of the face. However, it is observed in the experiment that the cycle consistency loss is difficult to maintain the anime face identity information (such as skin color, hair color, etc.) for many pictures. In addition, the facial expression of anime conversion task still lacks a clean labeled public dataset. Therefore, the task of facial expression conversion of anime characters has not been well studied.

Inspired by previous work, this paper proposes a facial expression of anime transformation method based on decoder control and attention mask. This method takes the GAN as the framework, and its generator has three sub networks: control information mapping network, expression transformation network and attention mask generation network. The control information mapping network maps discrete expression labels into high-dimensional control signals through affine transformation and nonlinear transformation. The expression transformation network realizes multiple control information injection of image features through AdaIN in the decoder part, so as to refine the expression control and enhance the effect. The attention mask generation network helps the expression control focus on the expression related areas by generating the attention mask of the input image to avoid the influence of the irrelevant areas. The contributions of this paper can be summarized as follows:

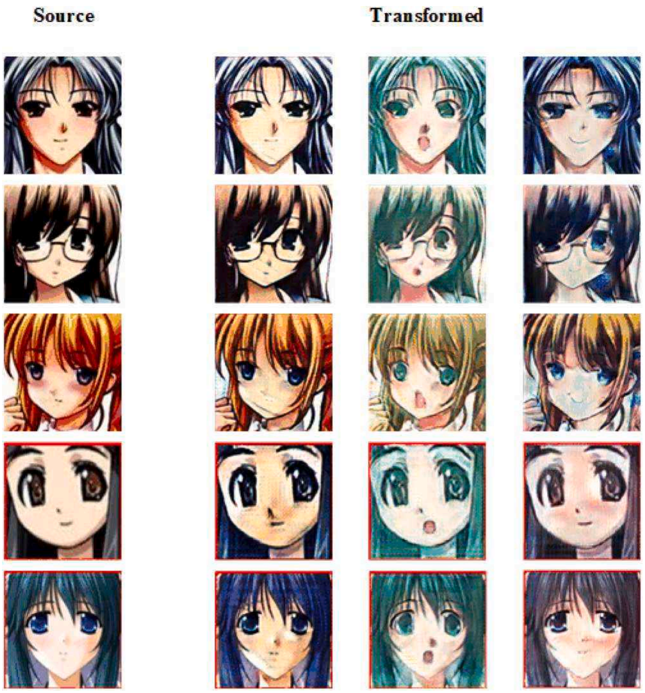

**Fig. 1.** Some failure examples by existing methods.

(1) A network generator based on decoder control is proposed to enhance the control over detailed transformation. The generator maps the discrete expression into high-dimensional control information by using a mapping network instead of simple copy operation, and then injects control information into different levels of features to enhance the control of expression features.

(2) The attention mask mechanism is integrated into the network proposed in this paper to help the expression control of the generator focus on the expression related areas, while retaining the characteristics of irrelevant areas, so as to better maintain the anime face identity information.

The rest of paper is organized as follows. Section 2 reviews the related works. Section 3 presents the framework of our proposed method. Details of our proposed method are described in Section 4. Section 5 shows the experimental results. Section 6 elaborates the conclusions.

## 2. Related work

### 2.1. Image-to-image translation

Image-to-image translation covers many applications situation in reality, which means it is one of the key directions in computer vision. The generative adversarial network proposed by Goodfellow et al. is an unsupervised learning technology, which has been widely used in image generation. The core idea is to establish two competing deep neural networks, which are trained simultaneously in a min max game. These years, GAN has achieved surprising success in the image generation task. Since the introduction of GAN, the image translation task has been greatly improved.

To solve the problem of style diversity, a lot of research work has been done on image translation [6–12]. Pix2Pix [6] is a popular supervised image translation framework that describes paired image translation as a general conditional GAN problem. It not only learns the mapping between input and output images, but also learns a loss function to train the mapping. To solve the problem that paired training data cannot be obtained, CycleGAN [7] learned the mapping from unpaired input fields to output fields by combining the adversarial loss with the cycle consistency loss. The core idea is that the generator network attempts to reconstruct the original image from the fake image and calculates the L1 norm as a cycle consistency loss. However, because these methods only focus on the mapping between the two domains, they are difficult to cope with the increasing number of domains. For example, if you have N domains, these methods require training (N-1) generators to handle the conversion between each domain, meanwhile the training time and the overhead of model parameters limit the actual use of these methods.

In order to solve the problem of domain scalability, some studies have proposed a unified framework [13–17]. Anoosheh et al. [14] proposed a multicomponent image translation model and training scheme, which is linearly related to the number of domains in terms of resource consumption and time required. The model proposed by Hui [15] consists of a globally shared auto-encoder and N Domain-Specific encoders/decoders, assuming that globally shared auto-encoders can map inputs to a common shared potential space. StarGAN [13] is one of the most famous research projects. It solves this problem by adding an auxiliary domain classifier to the discriminator network. Instead of training multiple times in the cross-domain model, it uses a single generator network to efficiently learn the mapping between different domains. The author realizes the expression migration of human face by treating each discrete expression as a domain. However, StarGAN still has some limitations, can hardly learn minor features. For instance, StarGAN is only able to learn major facial features such as hair color and skin color, while minor features like nose size or mustache are not effectively learned. In terms of facial expression conversion, StarGan's conversion effect is unsatisfactory, and it is challenging to preserve the original features. To address the limitation of StarGAN when dealing with minor features, we propose an expression transformation method with a mapping network, which aims to map discrete emoticon tags to high-dimensional control information, and then injected the information into the network multiple times during transformation. In addition, an integrated attention masking mechanism is introduced, allowing the expressive control of the network to focus more on expression-related features, improving the retention of the original information.

### *2.2. Anime content generation*

In the field of anime character image generation, Xiang et al.'s research [18] is able to generate anime portraits with fixed content and multiple styles from different artists. PSGAN [19] generates high-resolution full-body anime character images by gradually increasing the resolution of the generated images during training. Jin Y et al. [20] proposed a conditional anime face generation framework that can randomly generate anime faces with specific attributes. They collected an anime face dataset by crawling images from anime websites, automatically extracted 34 tags using illustration2Vec [21], and performed post-processing with SRGAN [22] to achieve super-resolution imaging. However, their generated images suffer from facial distortion and are difficult to identify. Li and Han's [23] SGA-GAN is a gender-style-based GAN that generates anime faces by adjusting gender attributes and style features. However, the generated anime faces have low quality and many distorted facial features.

In terms of facial attribute control of anime characters, Jiale Zhang et al. [24] proposed a cascaded pose transform network that unifies face pose transformation and head pose transformation, which generates anime character heads that can imitate speech actions from an anime-style image. They introduced a mask generator to make facial expression anime (for example, close eyes and open mouth). Honglun Zhang et al. [25] proposed a new generative adversarial network RAG. They suggested using a generator to learn residual attributes rather than target attributes to edit facial attributes. In the experiment, the author realized the editing of facial attributes (hair color, pupil color, mouth shape) of anime characters. Bing Li et al. [26] propose a novel generator architecture that can simultaneously transfer color/texture styles and transform local facial shapes into anime-like counterparts based on the style of a reference anime-face. Majid Mobini et al. [27] produced a dataset specially used for facial expression of anime conversion. Based on the unified framework StarGAN of multi domain image to image translation, adjusting the weight of the deceptive field of the generator network and the image reconstruction loss coefficient, a good facial expression conversion of anime is achieved. But its expression conversion effect is weak in some pictures, moreover there are cases where the face identity information of the source image cannot be retained. The problem exists partly due to the lack of public anime datasets containing expression tags, which has led to the fact that the task of facial expression conversion of anime characters has not been well studied. To address this problem, we propose a new facial dataset of anime characters, and the images in this dataset have been categorized based on the type of expression, which will be more helpful for subsequent work.

Though some researches have been done for facial expression transformation and made significant progress with the popularization of deep learning, many problems still need to be addressed. In summary, there are two main challenges in transforming expressions for anime images. Firstly, existing methods do not work well on anime datasets, resulting in a weaker control effect compared to real-image datasets. Secondly, it is challenging to maintain the original identity of the anime face while using existing methods.

## 3. Overview of method

The overall structure of our proposed method is illustrated in Fig. 2. Our method is based on GAN, composed by a generator and a discriminator. A pretrained network is introduced to calculate the contextual loss. Specifically, the task of the method is to use an image and a target expression label as inputs of the generator, so that the expression of the output image matches the target expression, while preserving other original attributes of the animated face, including identity, texture and posture information. In addition, in order for the generator to learn how to generate new images more stably, target expression label will be randomly generated. It is worth noting that during the training of the discriminator, the classification discriminant branch of the discriminator outputs a probability distribution describing the features of the input image, and this probability distribution will help the model to learn various feature attributes and classify them into the correct feature labels.

As illustrated in Fig. 3, the generator of our proposed method is composed by 3 sub-networks which are control information mapping network, expression transformation network and attention mask generation network. Control information mapping network can extend the numerically single discontinuous one-hot vector into numerically multivariate high-dimensional potential control information, so that the control information can refine the control of expression features. The structure of expression transformation network is similar to the encoder and decoder structure. This is because by separating the two phases of encoding and decoding, this architecture enables a flexible mapping from input to output. This mapping capability allows the model to handle more complex tasks. The encoder is composed of three down-sampling layers and three bottleneck layers meanwhile the decoder consists of three bottleneck layers and three up-sampling layers. Inspired by StyleGAN [28], we decided to control the image generation in the decoder part. The decoder control provides a better effect over the high-level attributes of the generated image, such as hair style, freckles, etc., than the input-side control and the feature-layer control. In order to achieve decoder control to enhance the effect of expression control, all but the last layer, the decoder part normalizes the image features in the way of AdaIN after each convolutional layer, so that the control information is injected many times in the process of image generation. The task of expression transformation network is to take an origin image and dimensional control information as inputs and output the expression region coloring mask matching the target expression. The structure of attention mask generation network is similar to expression transformation network. Origin image is taken as the input of the network. Then the attention mask with the same resolution as the image is outputted. Attentional masks can help to focus the model's control on the parts of the face that are more necessary for the transformation task,

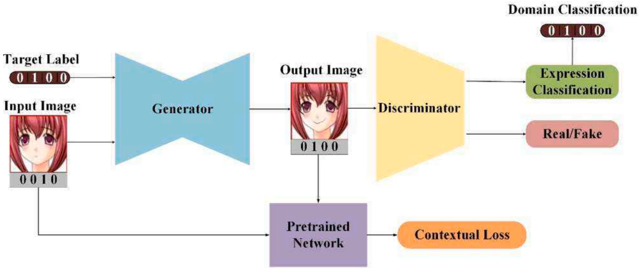

**Fig. 2.** The overall structure of our proposed method.

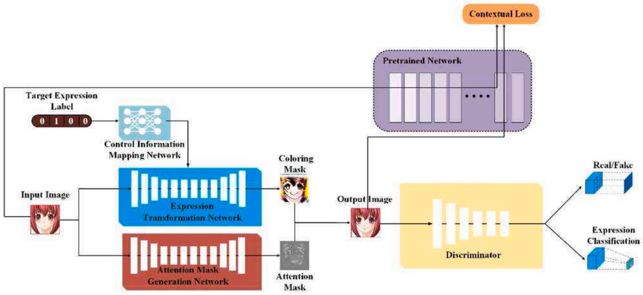

**Fig. 3.** The framework of our proposed method.

while the more minor parts are better preserved.

The discriminator of our proposed method consists of a shared convolutional layer and two subsequent branches. The shared convolutional layer is a common down-sampling structure, but no normalization operation is used in the feature extraction process. One of the branches of the discriminator is to judge the authenticity of generated image. The other branch is to classify the expression of generated image, which can enhance the transformation of expression.

The pretrained network calculate the contextual loss by comparing the features of fake and real images. Through minimizing the contextual loss, the generated output image will be similar to the input image, but not a perfect match, in order to preserve the identity information of the anime face from the input image while transforming the expression.

## 4. Details of method

### 4.1. Decoder control

Decoder control is achieved in the generator. For an anime face image $x$, label $c^o$ represents the original expression of image $x$, and label $c^t$ represents the target expression to be transformed. The task of the generator is to take $x$ and $c^t$ as inputs to make the expression of the output picture $y^*$ match the expression represented by $c^t$, while retaining other original attributes of the anime face, including identity, texture, pose and other information. In addition, in order for the generator to learn how to generate a new image more stably, the expression tag $c^t$ will be generated randomly. The definition of generator is shown in formula (1).

$$G(x,\ c^t) = y^* \tag{1}$$

Generator $G$ has three sub networks, as shown in Fig. 4, which are expression transformation network $G_{trans}$, attention mask generation network $G_{att}$ and control information mapping network $F$.

By learning the ability of affine transformation and nonlinear transformation of discrete expression labels in training, the control mapping network $F$ extends the numerically single discontinuous one-hot vector into numerically multivariate high-dimensional potential control information. This higher dimensional control information will contain more and deeper features during the training process, which will allow the model to better learn the differences of expressions and thus refine the expression features more. Let the input of the mapping network be discrete expression labels $c^t$ and the output be high-dimensional control information $k^t$. The definition of the relationship between them is shown in Formula (2).

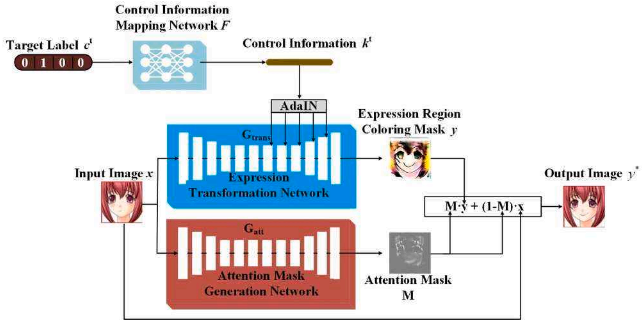
Control Information Mapping Network F
Target Label $c^t$
0 1 0 0
Control Information $k^t$
AdaIN
$G_{trans}$
Expression Transformation Network
Expression Region Coloring Mask $y$
Input Image $x$
Output Image $y^*$
M·y + (1-M)·x
$G_{att}$
Attention Mask Generation Network
Attention Mask M

**Fig. 4.** The structure of generator.

$$0k^t = F(c^t) \tag{2}$$

The mapping network is used to perform affine transformation and nonlinear transformation on the discrete expression labels, expanding them into high-dimensional control information while diversifying the values of the control information, avoiding the problems of a single value and redundancy of control information that exist in the above methods, and enriching the high-dimensional information with details of the expression control.

The structure of the control mapping network is a multi-layer perceptron, as shown in Fig. 5.

The multilayer perceptron has a strong nonlinear fitting capability, and through the superposition of multiple hidden layers and nonlinear transformation of the activation function, the multilayer perceptron is able to learn more complex feature representations, thus enabling the modelling of nonlinear relationships. The multi-layer perceptron consists of 9 fully connected layers and an activation layer after each fully connected layer. Each fully connected layer contains several perceptron models, also known as neurons. Every neuron contains several inputs and outputs. The relationship between input and output is linear, also known as affine transformation, which can be described as formula (3).

$$z = \sum\nolimits_i^m \omega_i x_i + b \tag{3}$$

where the $\omega_i$ stands for the weight of a neuron in layer $i$ and a neuron in layer $i + 1$, $b$ stands for the bias.

The first fully connected layer of the network is composed of 512 neurons, which maps k-dimensional discrete expression labels to 512 dimensions. Followed 7 fully connected layer is composed of 512 neurons, which is used to learn complex affine transformation. The last fully connected layer is composed of N dimensions, which maps the input into n-dimensional control information. Immediately following the output of each neuron is the activation layer, which is a nonlinear activation function, enabling the model to learn complex nonlinear transformations. The ReLU activation function is used here.

We achieve the decoder control in the expression transformation network $G_{trans}$, and multiple controls are realized by injecting control information into different levels of features in the decode, so as to enhance the control intensity of expression features in the network. $G_{trans}$ takes the image $x$ and control information $k^t$ as inputs and outputs

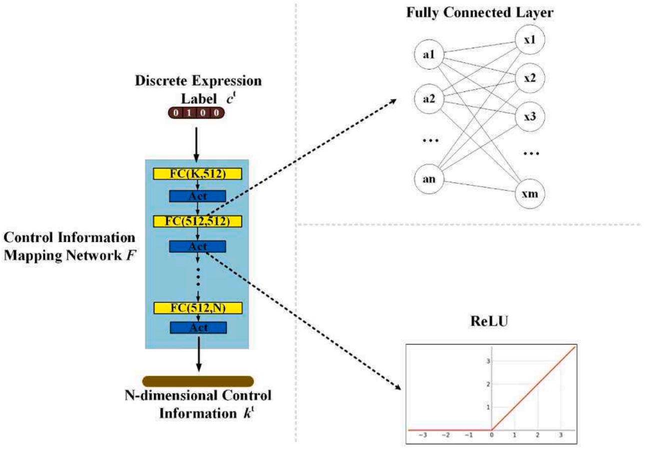
Fully Connected Layer
a1
a2
...
an
x1
x2
x3
...
xm
Discrete Expression Label $c^t$
0 1 0 0
FC(K,512)
Act
FC(512,512)
Act
FC(512,N)
Act
Control Information Mapping Network F
N-dimensional Control Information $k^t$
ReLU
3
2
1
−3
−2
−1
0
1
2
3

**Fig. 5.** The structure of the control mapping network.

the expression region coloring mask $y$ matching the target expression. The definition of the whole process is shown in Formula (4).

$$\mathrm{G}_{\mathrm{trans}}(x, k^t) = y \tag{4}$$

As shown in Fig. 6, The structure of $G_{trans}$ is similar to the encoder and decoder. K represents the size of convolution kernel, n represents the number of convolution kernel, and s represents the step size of convolution. The encoder is composed of three down-sampling layers and three bottleneck layers. The down-sampling layer uses convolution to extract image features, the bottleneck layer uses residual blocks to increase the depth of the network. After each convolutional layer, the IN layer is used for normalization. The decoder consists of three bottleneck layers and three up-sampling layers. The bottleneck layer also uses residual blocks to increase the depth of the network, and the up-sampling layer uses transpose convolution to enlarge the resolution of image features. In order to achieve decoder control to enhance the effect of expression control, all but the last layer, the decoder part normalizes the image features in the way of AdaIN [29] after each convolutional layer, so that the control information is injected many times in the process of image generation. Unlike existing multi-domain image-to-image translation methods that perform control information injection only once throughout the forward pass, the decoder control method we use injects control information using AdaIN in each subnetwork layer of the decoder or the latter part of the bottleneck layer of the generator as well as in the up-sampling layer, and the increase in the number of control times can play an enhanced role in the effect of the expression control, avoiding the expression problem of weak control strength. In more detail, firstly, the image features are normalized using the IN layer. Different from the IN layer of the encoder part, the affine transformation parameters here are set to be unlearnable. Then, a fully connected layer is used to affine transform the dimension of control information $k^t$ to twice the image feature dimension. Let the fully connected layer here be represented by $f$. Then slice the control information into two parameters $f(k^t)_1$ and $f(k^t)_2$ with the same dimension. Finally, $f(k^t)_1$ and $f(k^t)_2$ are used to affine transform the image features after IN layer to complete the injection of control information. Let the image feature be represented by $X$. The whole process of AdaIN can be represented by formula (5).

$$AdaIN(X, k^t) = \left(1 + f(k^t)_1\right)(IN(X)) + f(k^t)_2 \tag{5}$$

### 4.2. Attention mask

Attention masking is a technique used to control the flow of information by assigning exclusive weight maps to the output feature maps of the network layer to identify key features in the feature maps. These weight maps can be fed into the network as additional information to help the model focus on areas of the input information that need to be focused on. The principle of attention masking is that by introducing a new weight layer, the model can learn to automatically identify and focus on key information in complex input data, thereby improving the performance and accuracy of the model. Therefore, this paper constructs an attention mask generation network $G_{att}$ in the generator. As shown in Fig. 7, the structure of the network is composed of three down-sampling layers, six bottleneck layers and three up-sampling layers. In order to be able to extract deep features of the same dimension, the structure of $G_{att}$ is very similar to the expression transformation network $G_{trans}$, which is equivalent to replacing all AdaIN layers with IN layers, and changing the number of convolution kernels of the last layer to one. $G_{att}$ takes only the image $x$ as the input, and then outputs the attention mask $M$ with the same resolution as the image $x$ $(H \times W)$ but the number of channels is only 1. The definition of this process is shown in formula (6).

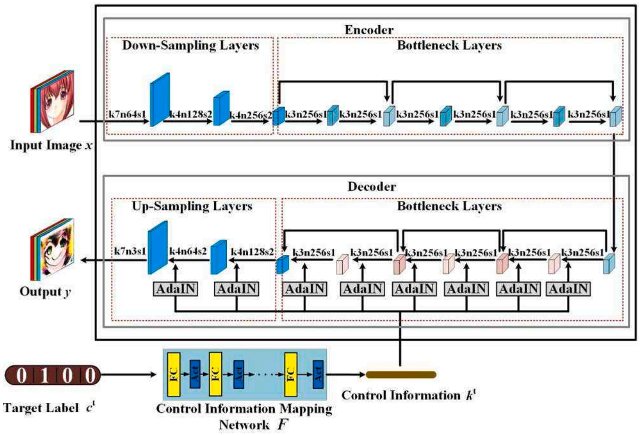

**Fig. 6.** The structure of expression transformation network.

$$M = G_{att}(x) \in \{0, \dots, 1\}^{H \times W} \tag{6}$$

Ideally, the value of $M$ should be 0 or 1, indicating that the corresponding pixel of $x$ is independent or related to the expression attribute. If an area is considered to be completely irrelevant to the transformation task (i.e., when $M = 0$), the model will not make any changes to that area, which will help the model to better preserve the original features of the image. Based on the expression region coloring mask $y$ output by $G_{trans}$ and the attention mask $M$ output by $G_{att}$, the identity information of the anime face can be maintained by extracting the changes of the expression related region only from $y$ and copying the other regions from $x$. This mechanism enables $G_{trans}$ to learn how to control the image features only in the expression related areas, while keeping the features unchanged in other areas to reduce the change of anime face identity information. The process of synthesizing the output image $y^*$ using the input image $x$, the expression region coloring mask $y$ and the attention mask $M$ can be described by formula (7).

$$y^* = M \cdot y + (1 - M) \cdot x \tag{7}$$

Therefore, the whole process of the generator can be described as formula (8).

$$G(x, c^t) = G_{att}(x) \cdot G_{trans}(x, F(c^t)) + (1 - G_{att}(x)) \cdot x \tag{8}$$

### 4.3. Discriminator and loss function

The discriminator of our proposed method consists of a shared convolutional layer and two subsequent branches, as shown in Fig. 8. The shared convolutional layer is a common down-sampling structure. One branch uses the structure of PatchGAN [6] to improve the authenticity of the generated image, the other branch improves the effect of expression feature conversion by capturing and determining the expression features of the generated image.

Like all GAN, in order to make the generated image indistinguishable from the real image, this paper uses adversarial loss, which is defined as formula (9). $D_{adv}(x)$ represents the image probability distribution of the branch output of judging whether the image is true or false in the

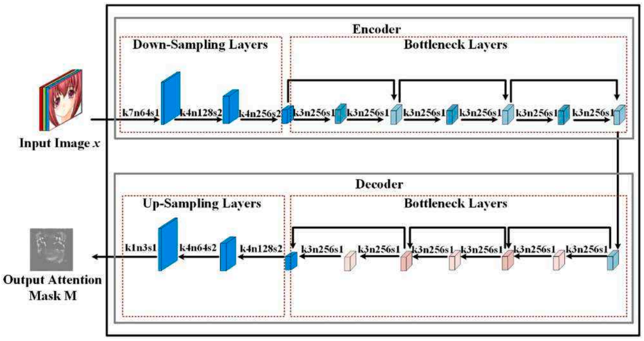

**Fig. 7.** The structure of attention mask generation network.

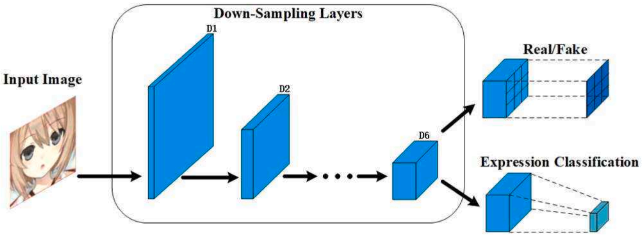

**Fig. 8.** structure of discriminator.

discriminator $D$. Generator $G$ attempts to minimize the target, while discriminator $D$ attempts to maximize the target.

$$L_{adv} = \mathbb{E}_x[\log D_{adv}(x)] + \mathbb{E}_{x,y}[\log(1 - D_{adv}(G(x,c^t)))] \tag{9}$$

At the same time, we apply WGAN-GP to stabilize the training process and generate higher quality images. As Ishaan et al. [30] explained, the gradient penalty is defined as formula (10), where $\widetilde{x}$ is the linear uniform sampling between the real sample and the generated sample pair.

$$L_{gp} = \mathbb{E}_{\widetilde{x}}\left[(\|\nabla_{\widetilde{x}} D_{adv}(\widetilde{x})\|_2 - 1)^2\right] \tag{10}$$

In order to make the expression in the generated image $y$ of generator $G$ match the target expression represented by $c^t$ as much as possible, this paper uses expression classification loss and divides it into two stages during training. In one stage, the expression loss is used to evaluate the expression category of the real image while training $D$, and in the other stage, the expression loss is used to evaluate the expression category of the generated image while training $G$. The definition of the former is shown in formula (11), where $D_{cls}(c^o|x)$ represents the expression probability distribution output by the expression classification branch in the discriminator $D$. $D$ learns to correctly classify the real image $x$ into its corresponding expression category $c^o$ by minimizing the loss.

$$L_{cls}^{real} = \mathbb{E}_{x,c^o}[-\log D_{cls}(c^o|x)] \tag{11}$$

On the other hand, the definition of the latter is shown in formula (12), where the generator $G$ learns how to generate a new picture conforming to the target expression category $c^t$ by minimizing the loss.

$$L_{cls}^{fake} = \mathbb{E}_{x,c^t}[-\log D_{cls}(c^t|G(x,\ c^t))] \tag{12}$$

In order to make the new image $y$ generated by the generator only change the expression attributes and retain other original attributes of the original image $x$ as much as possible, this paper applies contextual loss [31] to the generator.

Because the contextual loss requires that the output picture is similar to the original picture, it is allowed to produce position deformation. Due to the change of expression, the output picture and the input picture will not be fully aligned. Therefore, this deformation is advantageous. Compared with the cycle consistency loss and pixel by pixel matching, contextual loss is a good choice to maintain the output image anime face identity information. In this paper, a VGG19 [32] network trained in image classification is utilized as the loss network $\Phi$, which has learned to extract image features. Let the contextual loss be $L_{CX}$. In order for $G(x)$ to retain the same face identity information as $x$, the contextual loss needs to be calculated for the layer $l_t$ feature set extracted by $\Phi$. The definition of $L_{CX}$ is shown in formula (13).

$$L_{cx} = L_{CX}(G(x),\ x, l_t) \tag{13}$$

Finally, the comprehensive loss of training the generative adversarial network generator and discriminator as shown in formulas (14) and (15), where $\lambda_{cx}$ and $\lambda_{cls}$ are adjustable weight coefficients to control the weight of contextual loss and expression classification loss in the comprehensive loss respectively.

$$L_G = L_{adv} + \lambda_{cx}L_{cx} + \lambda_{cls}L_{cls}^{fake} \tag{14}$$

$$L_D = -L_{adv} + \lambda_{cls}L_{cls}^{real} + \lambda_{gp}L_{gp} \tag{15}$$

## 5. Experiment

### 5.1. Baseline

Since anime facial expression transformation can be considered as a specific application of unpaired multidomain image-to-image translation, in order to highlight the good performance of our proposed models, we selected CycleGAN [7], a dual-model based multidomain image-to-image translation framework, ComboGAN [14], a multi-model based unified multidomain image-to-image translation framework, and StarGAN [13], a single-model based multidomain image to image unified translation framework are compared. All of them are based on reconstruction loss approach to maintain face identity information. StarGAN is chosen as the baseline for the ablation experiment.

### 5.2. Dataset

Existing public anime face datasets such as iCartoonFace [33] provide high-quality anime face images, but the tags provided by these datasets are identity tags for anime face recognition tasks, moreover these tag data are not available for anime face expression tasks. For the lack of anime face dataset with expression label, this paper decides to create a dataset suitable for anime face expression task. Danbooru is a free image hosting website. Users can upload their tags and images, so each image carries the user's tag. In this paper, "angry", "sad", "smile" and "surprised" are selected as the basic expressions. The color anime pictures labeled by the four types of tags are retrieved and downloaded from the website. The downloaded pictures are trimmed by using the anime face detection model. Afterwards, we manually complete some pictures that are not normally cropped and marked, also eliminate some negative samples. Finally, this paper completes the data collection of four kinds of expressions (anger, sadness, smile and surprise). Fig. 9 shows some contents; Among them, 1615 pictures were collected in the angry expression dataset, 1138 pictures were collected in the crying expression dataset, 3370 pictures were collected in the smiling expression dataset, and 1183 pictures were collected in the surprised expression dataset. Lastly, the dataset is expanded to 11,612 pictures by using the data expansion method.

### 5.3. Training details

The experiment was carried out under the Pytorch deep learning framework. The GPU was Quadro RTX 4000 and the memory was 16 g (NVIDIA, Santa Clara, CA, USA). The model in this paper is trained using

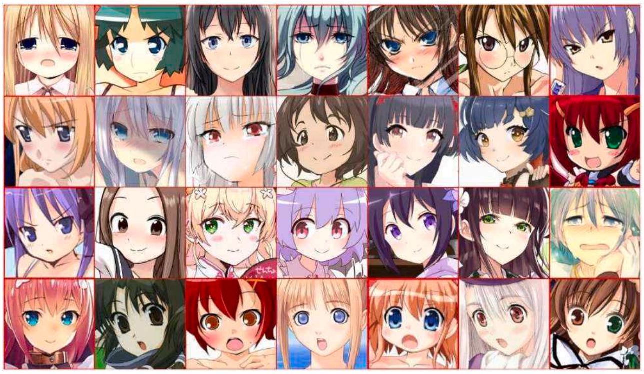

**Fig. 9.** partial image of dataset.

batch size (16), number of iterations (200,000) and optimizer ($\beta_1 = 0.5$ and $\beta_2 = 0.999$, Adam). The basic learning rate in the model is set to 0.0001. The basic learning rate is used in the first 100,000 iterations, in addition the quotient of the learning rate and the number of iterations is reduced every 1000 iterations between 100,000 and 200,000 iterations. The decreasing learning rate helps the model to gradually reduce the overfitting to the training data during the training process, thus improving the generalization performance of the model. $\lambda_{cx}$ and $\lambda_{cls}$ is set to 0.1 and 1. CycleGAN, ComboGAN and StarGAN use the default parameters provided by their authors for training.

### *5.4. Ablation experiment*

For the quantitative evaluation of the image quality generated, we compare the Fréchet Inception Distance (FID) [34] score and Kernel Inception Distance (KID) [35] score between the test image and the image generated by each method. The FID score and KID score are widely used metric for assessing the quality of generated image and have been shown to correlate well with human perception of image quality, making them a reliable and objective measure of image fidelity. The lower the FID score and KID score, the higher the quality of the generated image. The results are shown in Table 1.

The StarGAN is selected as a baseline of the experiments. Additionally, three baseline are considered in order to demonstrate that the improvement of our method has not resulted from other modifications than the proposed framework: First, baseline with decoder control (DC) demonstrates that the decoder control proposed can effectively enhance the quality of the transformation. Second, baseline with attention mask (AM) shows that the introduction of an attention mask is beneficial. Third, our method without contextual loss (CL) demonstrates that combining the decoder control and attention mask mechanisms simultaneously yields better results. Besides, the usefulness of the contextual loss has been proven. Our method improves the statistics of the FID metrics by 34.12 % compared to StarGAN in terms of expression transformation effect. And for the KID metric, which is closer to human intuitive perception, our method is 52.55 % ahead.

Additionally, Fig. 10 showcases the qualitative evaluation of different baselines. The results indicate that DC can markedly enhance the control effect of labels, particularly regarding the bending of the lips and changes in the eyes. However, injecting control information may render the generated images unstable in terms of details such as skin color and background. The problem causes the FID score of StarGAN+DC is worse than baseline. The use of AM solves the problem and improves the ability to retain detailed information like skin and hair color as well as the background. Our proposed method outperforms other baselines noticeably.

### *5.5. Quantitative evaluation*

For the quantitative evaluation of the generated image quality, we used the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) scores, which are commonly used to assess the quality of synthetic images in image translation research. FID measures the similarity of feature distributions between real and generated images, while KID calculates the statistical distance between two distributions using kernel methods. A lower FID or KID score indicates that the distribution of generated images is more similar to that of real anime-faces. By computing both metrics, we obtained an objective measurement of the quality of the generated images. The results are shown in Table 2.

The table shows that the FID and KID scores of the proposed method in this paper are significantly lower than those of other methods. On the FID metric, our method is 78.04 % ahead compared to the classical CycleGAN and 34.12 % ahead compared to StarGAN. And on the KID metric, our method is 51.31 % ahead of CycleGAN, and 52.55 % ahead compared to StarGAN. ComboGAN performs the worst in the experiment, and is significantly weaker than the rest of the control group on both metrics. This indicates that the quality of the images generated by the proposed method is higher than that of other methods, and that the original identity information is better preserved when transforming facial expressions.

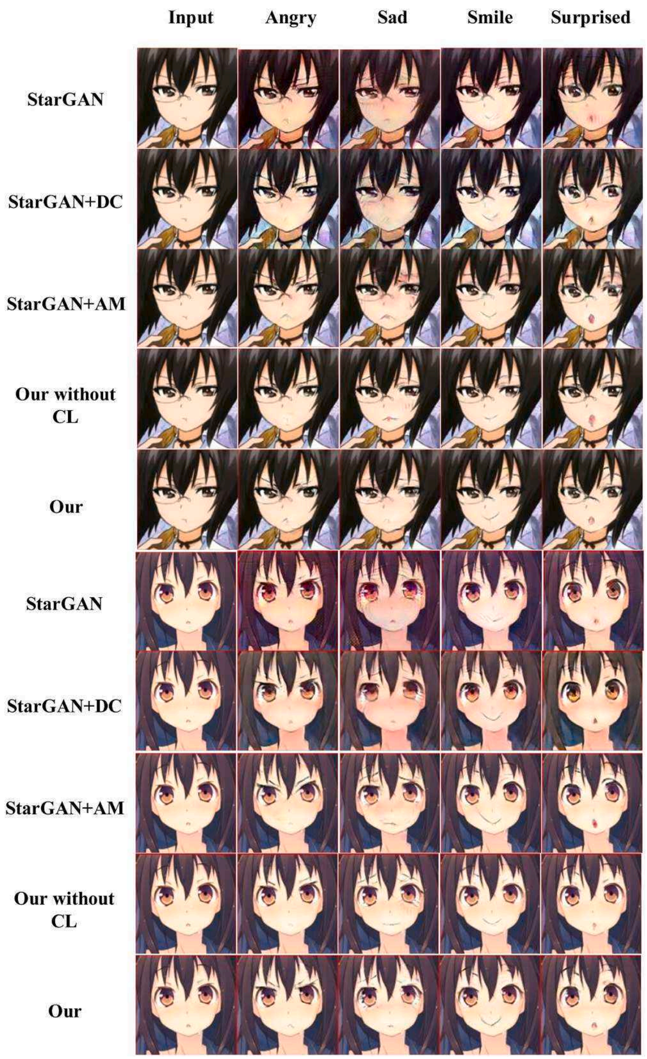

**Fig. 10.** partial results compared with other baselines.

**Table 1**
Fréchet Inception Distance (FID) scores of our method and baselines.

| Method | FID score | KID score |
|---|---|---|
| StarGAN [13] | 79.04 | 2.74E-02 |
| StarGAN+DC | 84.58 | 3.01E-02 |
| StarGAN+AM | 63.65 | 1.89E-02 |
| Our without CL | 54.49 | 1.41E-02 |
| **Our** | **52.07** | **1.30E-02** |

**Table 2**
comparison of FID scores with other methods.

| Method | FID score | KID score |
|---|---|---|
| CycleGAN [7] | 78.70 | 2.67E-02 |
| ComboGAN [14] | 141.98 | 7.96E-02 |
| StarGAN [13] | 79.04 | 2.74E-02 |
| Our | **52.07** | **1.30E-02** |

### 5.6. Qualitative evaluation

For the qualitative evaluation of the images generated by each method, Fig. 11 shows some experimental results of each method.

In terms of expression conversion effect, the performance of CycleGAN is unstable. Only the expression of individual results has changed in the first and fifth lines, which can explain the FID score of CycleGAN is better than StarGAN and ComboGAN. In the results of StarGAN, there are more effective expression transitions than CycleGAN; The degree of expression transformation in ComboGAN is stronger than that in StarGAN, but there are exaggerated changes. The method in this paper is more natural in vision.

In terms of maintaining face identity information, the expression conversion results of CycleGAN, ComboGAN and StarGAN are inconsistent with the face identity information of the original picture anime, which is reflected in the change of hair color and skin color, as well as the loss of nose in the form of weak embellishment in anime. In particular, ComboGAN, although it is able to change the details of the face more efficiently and is stronger in terms of transformation effect compared to both StarGAN and CycleGAN, too many original features are lost during the transformation process. ComboGAN is not able to keep the hair color information intact in almost every transformation, which explains why ComboGAN, which has a stronger control effect, has the worst scores in both FID and KID metrics. The methods proposed in this paper are consistent with the anime face of the original picture in these aspects. Therefore, the method proposed in this paper is superior to other methods in expression conversion effect and maintaining anime face identity information.

## 6. Conclusion

This paper proposes a method for transforming anime-style facial expressions based on decoder control and attention mask. Through the mapping network we proposed, the discrete expression labels are converted into high-dimensional control information, which is then incorporated into the decoder of the expression transformation network. This approach allows the discrete expression tags to contain more feature information and more effectively guide the transformation process of expression features, resulting in an improved transformation effect when using discrete expression tags. Furthermore, the proposed method integrates an attention mask mechanism that focuses the expression control of the network on relevant features to avoid affecting irrelevant features. This results in better preservation of the anime face identity information, ultimately improving the quality of the generated images. Experimental results demonstrate that the proposed method outperforms existing methods for transforming anime-style facial expressions based on general image translation models.

Our research has achieved some results in anime face expression transformation, however, there is still a lot of room for improvement in this area. Considering the data-dependent nature of deep learning, a high-quality anime face dataset is the main obstacle for current research. Existing working anime face datasets rely on collecting from the Internet during construction and require additional post-processing to obtain further annotations. And the ability of the model to generalize to a variety of anime faces in practice depends on the richness and diversity of the anime face categories in the dataset used, as well as the comprehensiveness of the coverage. Therefore, a high-quality anime face dataset remains a focus for future work.

Besides, the same type of expression can have different degrees of expression, for example, crying can be sobbing, sniveling, bawling, etc. Our method does not perform well in generalizing to different levels of expression, which may need to be addressed in the future with more careful data labelling and more flexible fine-tuning. Therefore, future work can further explore the diversity aspect and flexibility of expression generation.

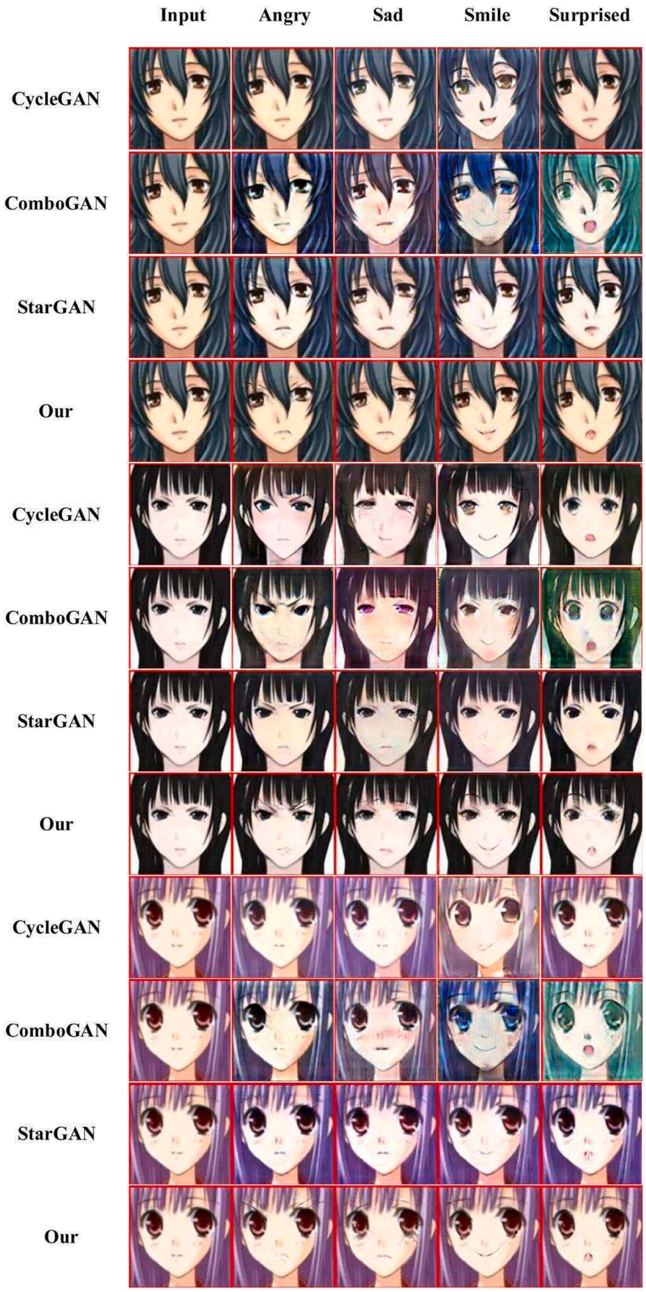

**Fig. 11.** partial results compared with other methods.

## CRediT authorship contribution statement

**Xinhao Rao:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Weidong Min:** Writing – review & editing, Conceptualization. **Ziyang Deng:** Supervision, Investigation. **Mengxue Liu:** Supervision, Investigation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No 62076117, the Jiangxi Provincial Key Laboratory of Virtual Reality under Grant No 2024SSY03151.

## Data availability

Data will be made available on request.

## References

[1] W. Min, R. Liu, D. He, et al., Traffic sign recognition based on semantic scene understanding and structural traffic sign location, IEEE Trans. Intell. Transp. Syst. 23 (9) (2022) 15794–15807.

[2] Q. Wang, W. Min, Q. Han, et al., Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification, Inf. Sci. (Ny) 564 (2021) 71–84.

[3] N. Cai, H. Chen, Y. Li, et al., Registration on DCE-MRI images via multi-domain image-to-image translation, Comput. Med. Imaging Graph. 104 (2023) 102169.

[4] Q.H. Cap, H. Uga, S. Kagiwada, et al., Leafgan: an effective data augmentation method for practical plant disease diagnosis, IEEE Trans. Autom. Sci. Eng. 19 (2) (2022) 1258–1267.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[6] P. Isola, J.Y. Zhu, T. Zhou, et al., Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[7] J.Y. Zhu, T. Park, P. Isola, et al., Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[8] K. Ko, T. Yeom, LeeM. SuperstarGAN, Generative adversarial networks for image-to-image translation in large-scale domains, Neural Netw. 162 (2023) 330–339.

[9] X. Li, S. Zhang, J. Hu, et al., Image-to-image translation via hierarchical style disentanglement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8639–8648.

[10] R. Komatsu, T. Gonsalves, Translation of real-world photographs into artistic images via conditional CycleGAN and StarGAN, SN Comput. Sci. 2 (6) (2021) 489.

[11] Q. Mao, H.Y. Lee, H.Y. Tseng, et al., Mode seeking generative adversarial networks for diverse image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1429–1437.

[12] X. Huang, M.Y. Liu, S. Belongie, et al., Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 172–189.

[13] Y. Choi, M. Choi, M. Kim, et al., Stargan: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[14] A. Anoosheh, E. Agustsson, R. Timofte, et al., Combogan: unrestrained scalability for image domain translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 783–790.

[15] L. Hui, X. Li, J. Chen, et al., Unsupervised multi-domain image translation with domain-specific encoders/decoders, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2044–2049.

[16] M.Y. Liu, X. Huang, A. Mallya, et al., Few-shot unsupervised image-to-image translation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10551–10560.

[17] Y. Xu, S. Xie, W. Wu, et al., Maximum spatial perturbation consistency for unpaired image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18311–18320.

[18] S. Xiang, H. Li, Anime style space exploration using metric learning and generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[19] K. Hamada, K. Tachibana, T. Li, et al., Full-body high-resolution anime generation with progressive structure-conditional generative adversarial networks, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, 0-0.

[20] Y. Jin, J. Zhang, M. Li, et al., Towards the high-quality anime characters generation with generative adversarial networks, in: Proceedings of the Machine Learning for Creativity and Design Workshop at NIPS, 2017.

[21] M. Saito, Y. Matsui, Illustration2vec: A Semantic Vector Representation of Illustrations[M]//SIGGRAPH Asia 2015, Technical Briefs, 2015, pp. 1–4.

[22] C. Ledig, L. Theis, F. Huszár, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[23] H. Li, T. Han, Towards diverse anime face generation: active label completion and style feature network, in: Eurographics (Short Papers), 2019, pp. 65–68.

[24] J. Zhang, C. Liu, K. Xian, et al., Large motion anime head anime using a cascade pose transform network, Pattern Recognit. 135 (2023) 109181.

[25] H. Zhang, W. Chen, J. Tian, et al., RAG: facial attribute editing by learning residual attributes, IEEE Access 7 (2019) 83266–83276.

[26] B. Li, Y. Zhu, Y. Wang, et al., Anigan: style-guided generative adversarial networks for unsupervised anime face generation, IEEE Trans. Multimed. 24 (2021) 4077–4091.

[27] M. Mobini, F. Ghaderi, StarGAN based facial expression transfer for anime characters, in: 2020 25th International Computer Conference, Computer Society of Iran (CSICC), IEEE, 2020, pp. 1–5.

[28] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks[C], in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[29] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

[30] I. Gulrajani, F. Ahmed, M. Arjovsky, et al., Improved training of wasserstein gans, Adv. Neural. Inf. Process. Syst. (2017) 30.

[31] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 768–783.

[32] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[33] Y. Zheng, Y. Zhao, M. Ren, et al., Cartoon face recognition: a benchmark dataset, in: The 28th ACM International Conference on Multimedia, 2020, pp. 2264–2272.

[34] M. Heusel, H. Ramsauer, T. Unterthiner, et al., Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. (2017) 30.

[35] M. Bińkowski, D.J. Sutherland, M. Arbel, et al., Demystifying mmd gans, in: Sixth International Conference on Learning Representations 6, 2018.