Heritage Science

RESEARCH

Open Access

Intelligent generation of Peking opera facial masks with deep learning frameworks

Ming Yan[1,2], Rui Xiong[1], Yinghua Shen[1], Cong Jin[1*] and Yan Wang[3*]

Abstract

The production of traditional Peking opera facial masks often relies on hand painting by experienced painters, which restricts the inheritance and development of this intangible cultural heritage. Current research mainly focuses on the digital reconstruction and storage of existing Peking opera facial masks, while high-quality facial mask generation technology is still in an infancy stage. In this paper, different deep learning frameworks are improved for learning features of Peking opera facial masks and generating new masks, which can effectively promote the creative application of Peking opera facial masks. First, using different data enhancement methods, an improved Style Generative Adversarial Network-2 (StyleGAN2) can learn implicit and explicit features of Peking opera facial masks and automatically generate new facial masks. In addition, an image translation framework for joint cross-domain communication under weak supervision is used to translate face sketches and color reference maps to an intermediate feature domain, and then synthesize new facial masks through an image generation network. The experimental results show that the generated Peking opera facial masks have good local randomness and excellent visual quality.

Keywords Peking opera facial masks, Image generation, Generative adversarial network (GAN), Image translation

Introduction

As one of the important performance aids in Peking opera, facial masks contain the unique cultural gene of Peking opera. Peking opera facial masks have rich colors, and their musical forms are complex and diverse. In the traditional production of Peking opera facial masks, it can only rely on hand painting by experienced painters. In the modern world of diverse entertainment, Peking opera facial masks, as an intangible cultural heritage, are becoming increasingly marginalized. In order to promote the inheritance and development of Peking opera facial masks, relevant studies have analyzed and expounded the symbolic meaning and aesthetics of Peking opera facial mask colors [1]. Integrated Peking opera facial mask elements into poster design to create modern graphics with traditional cultural elements [2]. Similarly, in order to inherit and develop the local tradition cultural, researchers have used augmented reality to create traditional masks with Indonesian cultural genes [3], and also used the Augmented Reality (AR) kit face tracking package to detect faces and implement an AR application for Malang masks [4].

The development and application of digital technology can bring technical possibilities for the preservation, transmission, and creative use of cultural heritage. For example, 3-Dimensional (3D) modeling techniques can realize interactive simulations of traditional costumes or Chinese landscape paintings [5, 6]. With the application of deep learning techniques for image generation, the effect of image generation can approach the level of human creation [7–9]. Applying deep learning

*Correspondence:
Cong Jin
jincong0623@cuc.edu.cn
Yan Wang
wy@cuc.edu.cn
[1] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China
[2] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Beijing 100024, China
[3] School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China

algorithms such as image generation and image translation to the generation of Peking opera facial masks can generate images of faces that inherit the relevant cultural genes and help preserve the culture of Peking opera.

However, there are still many challenges in the application of these deep learning frameworks to the generation of Peking opera facial masks. On the one hand, image generation techniques based on StyleGAN and its series of variants can generate realistic images by training on a large number of image datasets [7]. However, when the number of training datasets is insufficient, the discriminators of these networks are highly susceptible to over-fitting, which leads to severe instability in the training dynamics. Currently, there is no mature Peking opera facial mask dataset, which leads to the unsatisfactory generation of StyleGAN series network models based on a small number of Peking opera facial masks [7]. On the other hand, existing image translation models encode the style of the example image into a latent space and synthesize images that are similar to the style of the example images [10]. These approaches only consider the style of example images and fail to consider spatial correlations, leading in too poor network portability [11]. In addition, most of these image translation models applied to human faces and natural scenes are unable to learn well the hidden cultural genes in Peking opera facial masks, which results in most of the generated Peking opera facial masks with mixed lines and colors.

To solve the above problem, we use explicit data enhancement methods and a microscopic data enhancement method to improve the StyleGAN2 generation architecture, and adopt an advanced image translation framework. These deep learning frameworks can learn the features of Peking opera facial masks and generate new facial masks. In addition, to validate the performance of these deep learning frameworks, we construct a Peking opera facial mask dataset and conduct comprehensive experiments based on the dataset. Our contribution can be summarized as follows:

1. This paper delves into the benefits of different data enhancement approaches to model training. Different display data enhancement methods and implicit data enhancement methods are used to improve the network structure of StyleGAN2. It avoids discriminator over-fitting, trains good generators, and solves the drawback of poor generation for small datasets. The improved image generation architecture can effectively learn the explicit and implicit features of Peking opera facial masks and generate new Peking opera facial masks.
2. We use a cross-domain communication network for learning the correspondence between the input images. This image translation network translates face sketches and color reference maps into intermediate feature domains and establishes reliable dense correspondence. Then, the estimated feature correspondences are semantically aligned to the input through the image generation network to synthesize new Peking opera facial masks.
3. The original data are obtained through web download and e-book extraction. Then mirror transformation and other data enhancement methods are used to construct the Peking opera facial mask dataset which contains a total of 7128 pictures. Based on this dataset, a comparative analysis of the performance of the above two deep learning frameworks is performed. In order to further evaluate the quality of Peking opera facial mask generations, the generated images are used as test sets, and the relationship between the visual elements in Peking opera facial masks and the characters behind them are studied based on Residual Networks (ResNet) to complete the multi-classification task.

Related work

The development of deep learning provides a viable solution for people to use computers to generate images. Selecting a valuable image and converting it into another image with desired characteristics is a hot research topic in recent years, which is generally realized by image generation algorithms and image translation algorithms.

Image generation

The current mainstream image generation techniques include Variational Auto Encoder (VAE), Generative Adversarial Network (GAN), and Flow-Based Model (FBM). GANs which surpass other methods in image generation are considered a milestone in unsupervised learning [12]. It introduces the idea of game theory into the training process. The generator generates false images and the discriminator discriminates the data authenticity, and the final generated images are made to be false by the game between the two. In the same year, researchers proposed Conditional GANs (CGANs) with the addition of constraints [13]. CGANs introduce conditional variables into the generative and discriminative models, uses additional information to add conditions to the models, and guides the data generation process. StyleGAN2 [14] redesigns the normalization and designs a new network structure, which makes the network smoother, the hidden space more decoupled, and the quality of generated images higher. However, the realistic generation effect of StyleGAN2 benefits from a large number of images. In reality, it is still a challenge to collect a large image set for

a specific field. The key problem of training GANs model with a small data set is that the discriminator is easily over-fitted, which leads to serious instability of training dynamics. Data enhancement is a popular technique to alleviate the over-fitting of neural networks, especially for training data, so various data enhancement strategies are applied to the training of GANs. By introducing a random enhancement pipeline with 18 transforms, an Adaptive Data Augmentation (ADA) method is further designed to control the intensity of data enhancement adaptively [15]. Using pseudo-samples synthesized by the generator fed to a limited amount of real data, the Adaptive Pseudo Augmentation (APA) technique is presented to the discriminator in an adaptively enhanced manner [16]. The purpose of this pseudo-enhancement of real data is not to expand the real data set, but to inhibit the discriminator's confidence in distinguishing between real and spurious distributions. Diffusion-GAN randomly transforms data by using a differentiable forward diffusion process, which can be regarded as an enhancement method of domain agnosticism and model agnosticism [17].

Image translation

Image translation converts one representation of an image to another by understanding the mapping relationships between different image domains. CGAN is proposed to be applied to various supervised image translation tasks [18]. Pixel-level loss and GAN loss adopted by CGAN will lead to image blurring. When the image is severely distorted, CGAN and its variants are unable to capture the structural correspondence between different domains by a single translation network. To solve this problem, a multi-channel attentional selection GAN based on scene images and semantic mappings is proposed. Selection GAN can generate images of natural scenes from arbitrary viewpoints [9]. Subsequently, adaptive semantic consistency loss and style consistency discriminators are proposed to check whether image pairs are consistent in style [19]. The SPatially Adaptive DE-normalization (SPADE) method with an addition of a spatially adaptive normalization layer can further improve the quality of the generated images [20]. Essentially, the mapping from one image domain to another image domain is multi-modal. These basic models generally promote synthetic diversity by random sampling from a potential space. However, because the representation of the potential space is quite complex, none of these methods can provide fine control over the output or achieve good image-style correspondence. Unlike all the above approaches which only transfer global styles, an exemplar-based images translation framework Cross-domain Correspondence Network (CoCosNET) can transfer fine-grained styles from the semantic counterpart regions of the exemplars and enable the translation of images between different domains [21]. Therefore, the image translation framework meets the requirements of Peking opera facial mask generation and the generated Peking opera facial masks match the aesthetics of the public.

System model and algorithms

In this paper, we explore the generation method of Peking opera facial masks from two directions: image generation and image translation. For image generation, using the StyleGAN2 network which currently generates images with better quality, and avoiding discriminator over-fitting by using different data gain methods to improve the StyleGAN2. For image translation, CoCosNET, an image translation framework for cross-domain semantic communication, is used. It inputs semantic sketches and color style maps to provide more constraint information for the generated images.

Image generation network framework

In this paper, the image generation network adopts StyleGAN2, and different data gain methods are used to improve it. Next, we will first introduce the GAN network with the data enhancement module, and then detail the three data enhancement methods used in the experiments.

GANs with data enhancement module

General supervised learning and training tasks use data gain methods (rotation, color transformation, adding noise, etc.). Unlike classical data augmentation, the improved StyleGAN2 in this paper augments both true and false samples and lets the gradient propagate through the augmented samples to the generator. Specifically, the loss function of each network is formalized as follows:

$$L_D = -\mathrm{E}_{\mathbf{x}\sim p_\mathrm{r}}[g_1(D_\theta(T(\mathbf{x})))] - \mathrm{E}_{\mathbf{z}\sim p_\mathbf{z}}[g_2(D_\theta(T(G_\varphi(\mathbf{z}))))], \tag{1}$$

$$L_G = \mathrm{E}_{\mathbf{z}\sim p_\mathbf{z}}[g_2(D_\theta(T(G_\varphi(\mathbf{z}))))], \tag{2}$$

Where the generator G is a distribution mapping function that converts the low-dimensional potential distribution $p_\mathbf{z}$ to the target distribution p_g. The discriminator D evaluates the difference between the generated distribution p_g and the true distribution p_r. E$(\cdot)$ indicates the calculation of the average function. Furthermore, $\mathbf{z}$ representatives the random latent code and $\mathbf{x}$ representatives the input image. g_1 and g_2 stands for different functions for different GAN. φ and θ respectively stands for the parameter of generator and the parameter of discriminator. $T(\cdot)$ indicates the enhancement of different data

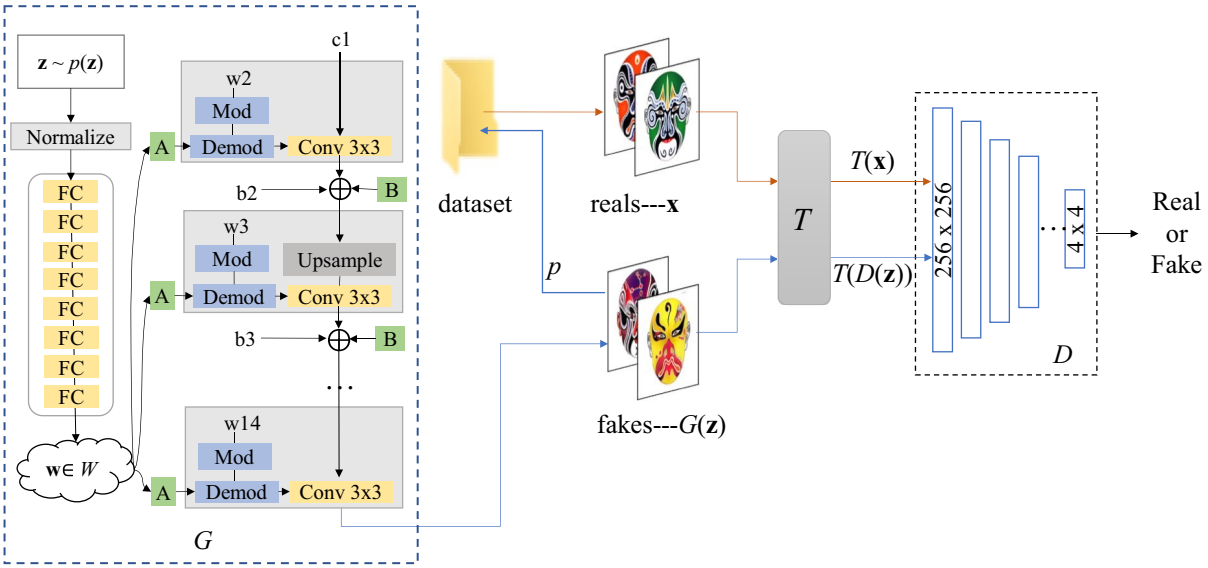

Fig. 1 Stylgan2 network structure with data expansion modules (*T* is the data expansion module.)

conversion technologies. Standard data conversion methods can usually be divided into two types. One is spatial transformation, including rotation, flip, translation, and geometric transformations such as scaling and stretching. The other is visual transformation, such as brightness and color. In practice, we also investigate for these two data enhancement methods. Figure 1 shows the framework of the GAN network after adding the data expansion modules, using the StyleGAN2 image generation model as an example. We can see that the input data of the discriminator is processed by the data expansion module.

If the data augmentation is used directly in the training task of GAN, the network after adding the data expansion module may lead to gain leakage to the generator. For example, if a rotation transformation is added to the data gain module, the generator will also generate rotated images. Experimental results of data perturbation on GANs have shown that if the transformation of the data is reversible in terms of the probability distribution, then the training of the model will find the correct distribution bypassing the data perturbation [22]. This shows that the training of the generator is not affected and the problem of gain leakage does not occur when the data gain is reversed. In our experiments, the strength of the data enhancement is defined as a scalar p, and controlling $p \in [0,1)$, then the data gain can become reversible.

Adaptive discriminator enhancement module

Since GANs training is in a dynamic equilibrium, dynamic control of the enhancement strength according to the degree of discriminator over-fitting can avoid the complexity and computational cost associated with manual adjustment. Therefore, in order to quantify over-fitting, we study a series of reasonable heuristics derived from the original output logic of the discriminator and then match the appropriate target value to automatically adjust p.

In the experiments of this paper, a total of three adaptive discriminator enhancement methods, ADA [15], APA [16], and Diffusion-GAN [17], are investigated. ADA is an adaptive discriminator enhancement mechanism that introduces a random enhancement pipeline with 18 transforms. APA uses generators to mitigate over-fitting and augment the true data distribution with the generated images. Diffusion-GAN uses a microscopic data enhancement approach to inject instance noise into the data with a Gaussian mixture distribution generated by a forward diffusion chain. The heuristic algorithms for each method are specified below.

The ADA defines the part of the heuristic algorithm used to estimate the training set to obtain the output of the positive discriminator. The over-fitting heuristic algorithm can be represented by the following equation:

$$rt = \mathrm{E}[\mathrm{sign}(D\mathrm{train})] \tag{3}$$

For this heuristic, $r_t = 0$ means no over-fitting and $r_t = 1$ means complete over-fitting. Dtrain denotes the discriminator output of the training set, sign(·) represents the symbolic function, and E[·] denotes the average of N consecutive small-batch discriminators. In practice, we use $\mathrm{N} = 4$ to correspond the images of $4 \times 64 = 256$.

The strategy for using *rt* to adjust p is as follows. Firstly, setting a threshold value t and initializing p to zero. If *rt* represents too much/too little over-fitting (i.e., greater

than/less than t) about t, the probability p will increase/decrease by a fixed step. The experimental setup adjusts p once every four small batches and clamps p from below to 0 after each adjustment. In this way, the intensity of data enhancement can be autonomously controlled according to the degree of over-fitting.

As GAN itself has powerful image generation capability, it is also a natural and feasible scheme for data enhancement to use the generator in GAN to generate images. APA solves the problem that ADA may lead to over-fitting of discriminator. It uses the standard data conversion method to process the generated data and takes it as a new data enhancement method. The new heuristic can be calculated by the following equation:

$$\begin{aligned} \lambda_t = & \ \mathrm{E}(\mathrm{sign}(D_{\mathrm{real}})) \\ = & \ \mathrm{E}(\mathrm{sign}(\mathrm{logit}(D(\mathbf{x})))) \end{aligned} \tag{4}$$

Note that the strategy of using λ_t to adjust p is consistent with the strategy of using rt to adjust p mentioned above and logit(·) stands for logit function. As the generated data is passed to the discriminator as real data for training, the optimization objective is updated with the following equation:

$$\begin{aligned} \min_G \max_D V(G,D) = & (1-\alpha)\mathrm{E}_{\mathbf{x}\sim p_{\mathrm{r}}(\mathbf{x})}[\log D_\theta(\mathbf{x})] \\ & + \alpha \mathrm{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log D_\theta(G_\varphi(\mathbf{z}))] \\ & + \mathrm{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1-D_\theta(G_\varphi(\mathbf{z})))] \end{aligned} \tag{5}$$

Where $V(D, G)$ represents the difference between the generation model and the discrimination model. α is the expected intensity that approximates the dynamic adjustment effect throughout the training process. Since $p \in [0,1)$, specifies $0 \le \alpha < p_{\max} < 1$, where $p_{\max}$ is the maximum value of the intensity of the pseudo-sample added throughout the training process.

Diffusion-GAN adds noise to the input data of the discriminator for the purpose of data enhancement. The network incorporates a diffusion model that injects noise from a Gaussian mixture distribution to achieve data transformation (The Gaussian mixture distribution consists of weighted diffusion samples from clean images at different time steps.). Diffusion-GAN min-max target is defined as following equation:

$$\begin{aligned} & \min_G \max_D V(G,D) \\ & = \mathrm{E}_{\mathbf{x}\sim p_{\mathrm{r}}(\mathbf{x}),t\sim p_\pi,\mathbf{y}\sim q(\mathbf{y}|\mathbf{x},t)}[\log D_\theta(\mathbf{y},t)] \\ & \quad + \mathrm{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),t\sim p_\pi,\mathbf{y}_{\mathrm{g}}\sim q(\mathbf{y}|G_\varphi(\mathbf{z}),t)}[\log(1-D_\theta(\mathbf{y}\mathrm{g},t))] \end{aligned} \tag{6}$$

For $\forall t \in \{1,\cdots,T\}$, the discriminator learns how to distinguish the diffusion-generated sample $\mathbf{y}_{\mathrm{g}}$ from the diffusion-true observation $\mathbf{y}$. The specific priority is determined by the value of the mixture weight π_t (non-negative and summing to 1). $\mathbf{y} \sim q(\mathbf{y}|\mathbf{x},t)$ stands for the edge distribution that becomes $\mathbf{x}$ after t steps of the forward diffusion chain. $\mathbf{y}\mathrm{g} \sim q(\mathbf{y}|G_\varphi(\mathbf{z}),t)$ can be re-parameterized as follows: $\mathbf{y}\mathrm{g} = \sqrt{\overline{at}}G_\theta(\mathbf{z}) + \sqrt{(1-\overline{at})}\sigma\varepsilon, \varepsilon \sim \mathrm{N}(0,I)$. Both the pre-defined variance schedule βt and variance σ^2 are defined by the forward diffusion chain, $at = 1-\beta t$, $\overline{at} = \prod_{i=1}^{t} ai$. The gradient which is calculated according to Eq. (6) can be directly back-propagated to the generator.

This paper delves into the benefits of different data enhancement approaches to model training. Explicit enhancement includes geometric transformations, color transformations, and data expansion using pseudo-samples generated by the generator. The micro-enhancement method is to randomly transform the data by using the micro-forward diffusion process and inject the instance noise. The StyleGAN2 improved by data augmentation can well avoid the discriminator over-fitting. It can enhance the stability of the training process, and improve the fidelity of the generated face images.

Image translation network framework

In this paper, the image translation network model uses a joint cross-domain communication framework under weak supervision. The Peking opera facial mask sketches belong to the source domain and the color reference images belong to the target domain. The framework implements the translation from the source domain to the target domain. Firstly, the semantic correlation between the source domain and the target domain is found. Then the sample images are distorted. Finally, new Peking opera facial masks are generated according to the distorted images. The line features of the generated images come from the source domain, while the color features are similar to the target domain. The algorithm framework for the image translation model proposed in this paper is shown in Fig. 2, which contains two parts: the cross-domain alignment network and the image composition network.

Cross-domain alignment network

Cross-domain alignment network realizes mapping the images of the source domain and the target domain to the intermediate domain W, and finds the semantic correspondence in the intermediate domain, thereby distorting the sample images. Usually, the semantic correspondence is obtained by matching the patch [23] in the feature domain with a pre-trained classification model. As the pre-training models are trained according to the images in specific scenes, they are only suitable

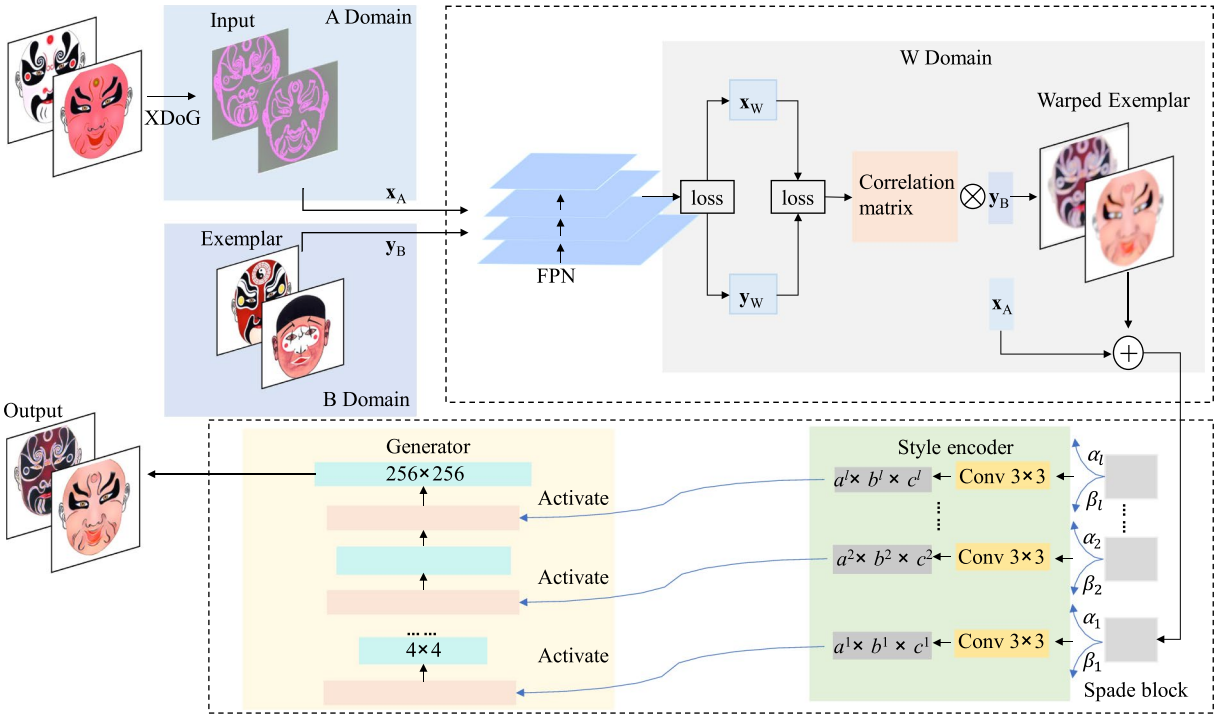

Fig. 2 Image translation network framework

for the semantic correspondence of specific images, so they cannot be used to express the semantic information of images in different scenes. The cross-domain alignment network model adopted in this paper can map the semantic information of different domains to the intermediate domain W and find the corresponding relationship in the W domain, which is suitable for most images. In the experiment of this paper, the model can accurately find the semantic correspondence information between the Peking opera facial masks' sketches and the color reference images.

The input sketches $\mathbf{x}_A$ are from domain A and the reference images $\mathbf{y}_B$ are from domain B. Firstly, $\mathbf{x}_A$ and $\mathbf{y}_B$ are fed into a Feature Pyramid Network (FPN) [24] to extract different scale feature maps, which are then converted into representations $\mathbf{x}_W$ and $\mathbf{y}_W$ in the W domain by a transformation relation. The conversion relationships are expressed as follows:

$$\mathbf{x}_W = F_{A\to w}(\mathbf{x}_A; \theta_F, A \to W), \tag{7}$$

$$\mathbf{y}_W = F_{B\to w}(\mathbf{y}_B; \theta_F, B \to W), \tag{8}$$

Where F denotes the transformation relationship from the input domain to the intermediate domain. $\mathbf{x}_W$ and $\mathbf{y}_W$ represent the semantic features of images in the intermediate domain. $\mathbf{x}_A$ and $\mathbf{y}_B$ denote the semantic features of the two input domains. θ is the parameter that need to be learned. Because the Peking opera facial masks' sketches and the color images are different domain images in the same scene and contain the same semantics, the conversion of $\mathbf{x}_A$ and $\mathbf{y}_B$ to W domain is completely aligned. The loss corresponding to this step is the loss of domain alignment:

$$L_{\text{domain}} = \|F_{A\to W}(\mathbf{x}_A) - F_{B\to W}(\mathbf{x}_B)\|, \tag{9}$$

After converting both the sketches in domain A and the references in domain B to domain W, the relationship between $\mathbf{x}_W$ and $\mathbf{y}_W$ is found. The similarity matrix between them is calculated according to Eq. (10), where each pair of elements $\mathbf{x}_W(u)$ and $\mathbf{y}_W(v)$ are characteristically related.

$$\mathbf{M}(u,v) = \frac{\hat{\mathbf{x}}_W(u)^T \hat{\mathbf{y}}_W(u)}{\left\|\hat{\mathbf{x}}_W(u)\right\| \left\|\hat{\mathbf{y}}_W(u)\right\|} \tag{10}$$

The matrix element $\mathbf{M}(u,v)$ represents the semantic similarity between $\mathbf{x}_W(u)$ and $\mathbf{y}_W(v)$. $\hat{\mathbf{x}}_W(u)$ and $\hat{\mathbf{y}}_W(v)$ denote the channel concentration characteristics of

elements $\mathbf{x}_{\mathrm{W}}(u)$ and $\mathbf{y}_{\mathrm{W}}(v)$, which are calculated by Eqs. (11–12):

$$\hat{\mathbf{x}}_{\mathrm{W}}(u) = \mathbf{x}_{\mathrm{W}}(u) - mean(\mathbf{x}_{\mathrm{W}}(u)), \tag{11}$$

$$\hat{\mathbf{y}}_{\mathrm{W}}(v) = \mathbf{y}_{\mathrm{W}}(v) - mean(\mathbf{y}_{\mathrm{W}}(v)), \tag{12}$$

By weighted summation of the pixels with the highest similarity in $\mathbf{y}_{\mathrm{B}}$, the corresponding relationship between domain B and domain A can be obtained, which is marked as $r_{\mathbf{y}\to\mathbf{x}}$. With proper learning, it is possible to obtain $\mathbf{y}_{\mathrm{B}}$ by transforming into A domain. In other words, it is possible to obtain an image of the sample images after distorting them according to the lines of the input sketch. A correspondence regularity term is introduced here. The regularity term requires that the original images are obtained by distorting the lines of the original sample images in the same way. The loss corresponding to this step is called the correspondence regularity loss:

$$L_{\mathrm{reg}} = \left\| r_{\mathbf{y}\to\mathbf{x}\to\mathbf{y}} - \mathbf{y}_{\mathrm{B}} \right\|, \tag{13}$$

where $r_{\mathbf{y}\to\mathbf{x}\to\mathbf{y}}$ is the correspondence from domain B to domain A and then to domain B.

Image composition network

The purpose of the image composition network is to progressively generate high-quality target domain images by using the input sketch $\mathrm{x_A}$ and the distorted example image $\mathrm{r_{y\to x}}$ aligned with it.

The image composition network mainly uses a modified spatially adaptive de-normalization layer. For style injection, a multi-layer convolution method is used to gradually inject style information of distorted images. In terms of structure generation, a SPADE block is used to project distorted images generated by the cross-domain alignment network to different activation locations in order to better preserve the image structure information synthesized by previous layers [20].

The simple transformation is used to map $\mathbf{x}_{\mathrm{A}}$ and $r_{\mathbf{y}\to\mathbf{x}}$ to modulation coefficients of the regularization layer, and then the style of the generated images is controlled by modulating the regularization layer. The transformation process is implemented by Eq. (14).

$$\alpha_{a,b}(r_{\mathbf{y}\to\mathbf{x}}, \mathbf{x}_A) \times \frac{F_{c,a,b} - \mu_{a,b}}{\sigma_{a,b}} + \beta_{a,b}(r_{\mathbf{y}\to\mathbf{x}}, \mathbf{x}_A) \tag{14}$$

where the conditional inputs $\mathbf{x}_{\mathrm{A}}$ and $r_{\mathbf{y}\to\mathbf{x}}$ stand for 2-Dimensional images. c is the channel dimension. a and b are the feature space sizes. $\sigma_{a,b}$ and $\mu_{a,b}$ stand for regularization statistics. $\alpha_{a,b}$ and $\beta_{a,b}$ stand for de-normalization coefficients.

To learn the correct correspondence, the image composition network imposes the following losses on output images: losses for pseudo exemplar pairs, perceptual loss, contextual loss, and adversarial loss.

Losses for pseudo exemplar pairs $\mathbf{x}_{\mathrm{A}}$ and $\mathbf{x}_{\mathrm{B}}$ are pairs of images semantically aligned in domain A and domain B. $\mathbf{x}_{\mathrm{B}}^{'}$ is the image of $\mathbf{x}_{\mathrm{B}}$ after geometric deformation. If $\mathbf{x}_{\mathrm{B}}^{'}$ is used as the example image and $\mathbf{x}_{\mathrm{A}}$ is used as the input image, $\mathbf{x}_{\mathrm{B}}$ is the generated image.

$$L_{\mathrm{feat}} = \sum_{l} \lambda_l \left\| \varphi_l(G(\mathbf{x}_{\mathrm{A}}, \mathbf{x}_{\mathrm{B}}^{'}) - \varphi_l(\mathbf{x}_{\mathrm{B}}) \right\|, \tag{15}$$

where φ_l represents the layer l activation of Visual Geometry Group (VGG-19) and λ_l is the equilibrium parameter.

Perceptual loss Penalizing perceptual loss to minimize semantic differences.

$$L_{\mathrm{perc}} = \left\| \varphi_l(\hat{\mathbf{x}}_{\mathrm{B}}) \to \varphi_l(\mathbf{x}_{\mathrm{B}}) \right\|, \tag{16}$$

where $\hat{\mathbf{x}}_{\mathrm{B}} = \mathbf{x}_{\mathrm{B}} - mean(\mathbf{x}_{\mathrm{B}})$, φ_l represents the activation of layer l of VGG-19.

Contextual loss To keep the output image in the same style as the example images, context loss is used to match the statistical distribution of $\mathbf{x}_{\mathrm{B}}^{'}$ and $\mathbf{y}_{\mathrm{B}}$ on the underlying feature map of VGG19.

$$L_{\mathrm{context}} = \sum_{l} \omega_l [-\log(\frac{1}{n_l} \sum_{i} \max_{j} \mathrm{A}^l(\phi_i^l(\hat{\mathbf{x}}_{\mathrm{B}}), \phi_j^l(\mathbf{y}_{\mathrm{B}})))], \tag{17}$$

where ω_l controls the relative importance of layer l. n_l represents the features contained in the l-th layer activation of VGG-19. A^l represents the l-th level of the A domain.

Adversarial loss Similar to the loss function of the general GANs, the main purpose of adversarial loss is to make the generated images belong to the B domain and improve the quality of the generated Peking opera facial masks. The discriminator D is trained alternately with the translation network G, which eventually makes the generated images difficult to distinguish from the example images.

$$L_{\mathrm{adv}}^{D} = -E[h(D(\mathbf{y}_{\mathrm{B}}))] - E[h(-D(G(\mathbf{x}_{\mathrm{A}}, \mathbf{y}_{\mathrm{B}})))], \tag{18}$$

$$L_{\mathrm{adv}}^{G} = -E[D(G(\mathbf{x}_{\mathrm{A}}, \mathbf{y}_{\mathrm{B}}))], \tag{19}$$

where $h(t)$ representatives the hinge function for the regularized discriminator.

The total loss of the image translation framework is the weighted sum of the loss of the cross-domain alignment network and the image generation network.

$$L = \min_{G} \max_{D} \varphi_1 L_{feat} + \varphi_2 L_{perc} + \varphi_3 L_{context} + \varphi_4 L_{adv} + \varphi_5 L_{domain} + \varphi_6 L_{reg}, \quad (20)$$

where $\varphi_1 \sim \varphi_6$ are the weights, which are used to balance several losses and generate high quality target images.

The image cross-domain communication network can accurately correspond to the semantic relationship between Peking opera facial masks' sketches and color references so that the generated images will have the characteristics of real Peking opera facial masks. The line relationship between the different colors of the generated images is clear, which can accurately correspond to the line characteristics of Peking opera facial masks' sketches and the color characteristics of reference images.

Experimental results and analysis

This section mainly includes several parts such as dataset creation, experimental environment introduction, experimental result analysis, and conclusion.

Datasets

Up to now, no publicly available datasets of Peking opera facial masks can be found. We download images from the web related to Peking opera facial masks and screenshots of e-books as the original images. Since the hand-drawn and computer-drawn images differ in color and detail, there is no duplication problem, so the two are finally combined to get a total of 1782 raw data. For the generation task, the more data required the better generation results, whether it is an image generation algorithm or an image translation algorithm. In order to avoid the problem of generator leakage caused by excessive data enhancement, the data enhancement operation is only completed by image inversion and color change of the dataset. Finally, we get 7128 Peking opera facial masks.

The line drawings needed for the image translation algorithm are obtained by extracting the face edges, which use the extended difference-of-Gaussians algorithm [25]. All images in the experiment are preprocessed to a uniform resolution of 256*256.

Experimental environment

The network models of the two methods adopted in this paper are based on the Pytorch framework. The running platform configuration CPU is an Intel Core i7-9700 K processor, which with 12 M cache, 3.60 GHz, 8 cores, and 16G RAM. The two GPU models are NVIDIA GeForce GTX 1080 Ti with 11 GB of dedicated memory.

Qualitative analysis

In terms of image generation, the Peking opera facial masks generated by StyleGAN2 and the model improved by three different data enhancement methods have good visual effects. Some of these images almost reach the quality of the original dataset images from the naked eye. Figure 3 shows the comparison of the visual quality of generated images by different methods. However, we can also find that some images generated by StyleGAN2 have uneven colors and interrupted lines (The images are highlighted by the red boxes in Fig. 3). In contrast, the data-enhanced model produces higher-quality images. The three data enhancement methods which are studied in this paper, including two display data enhancement (APA and ADA) and one microscopic data enhancement (Diffusion-GAN), can produce images that are difficult to distinguish between true and false.

In terms of image translation, the CoCosNET framework used in this paper generates images that combine the shape characteristics of line drawings and the color characteristics of reference drawings. The generated images also have good visual effects, as shown in Fig. 4. In addition, another advanced image translation method, SPADE, is used in this paper as a comparison. Figure 5 shows the comparison of the generated images by both SPADE and CoCoSNET methods. The colors and lines of SPADE generated images are messy. They only have the rough features of Peking opera facial masks and cannot be used as usable images. In contrast, the face images which are generated by CoCosNET achieve a level of quality similar to the real images. In addition, because the dataset used in this paper does not annotate the Peking opera facial masks in fine chunks, but directly extracts the overall line drawings of the images, the process of coloring will lead to color mixing in some areas of some images (The part marked with red boxes in Fig. 5.). Such an approach can form a new Peking opera facial mask style.

According to the evaluation of the Peking Opera Facial Masks Expert Group, the Peking opera facial masks generated by a series of improved StyleGAN2 models and CoCosNET models studied in this paper are in line with the traditional masks making standards in color and spectrum. This proves that the generated masks can be used in different aspects of the practice, such as mask design and Peking opera performance. This proves that the automatic generation technology of Peking opera facial masks studied in this paper can provide new ideas

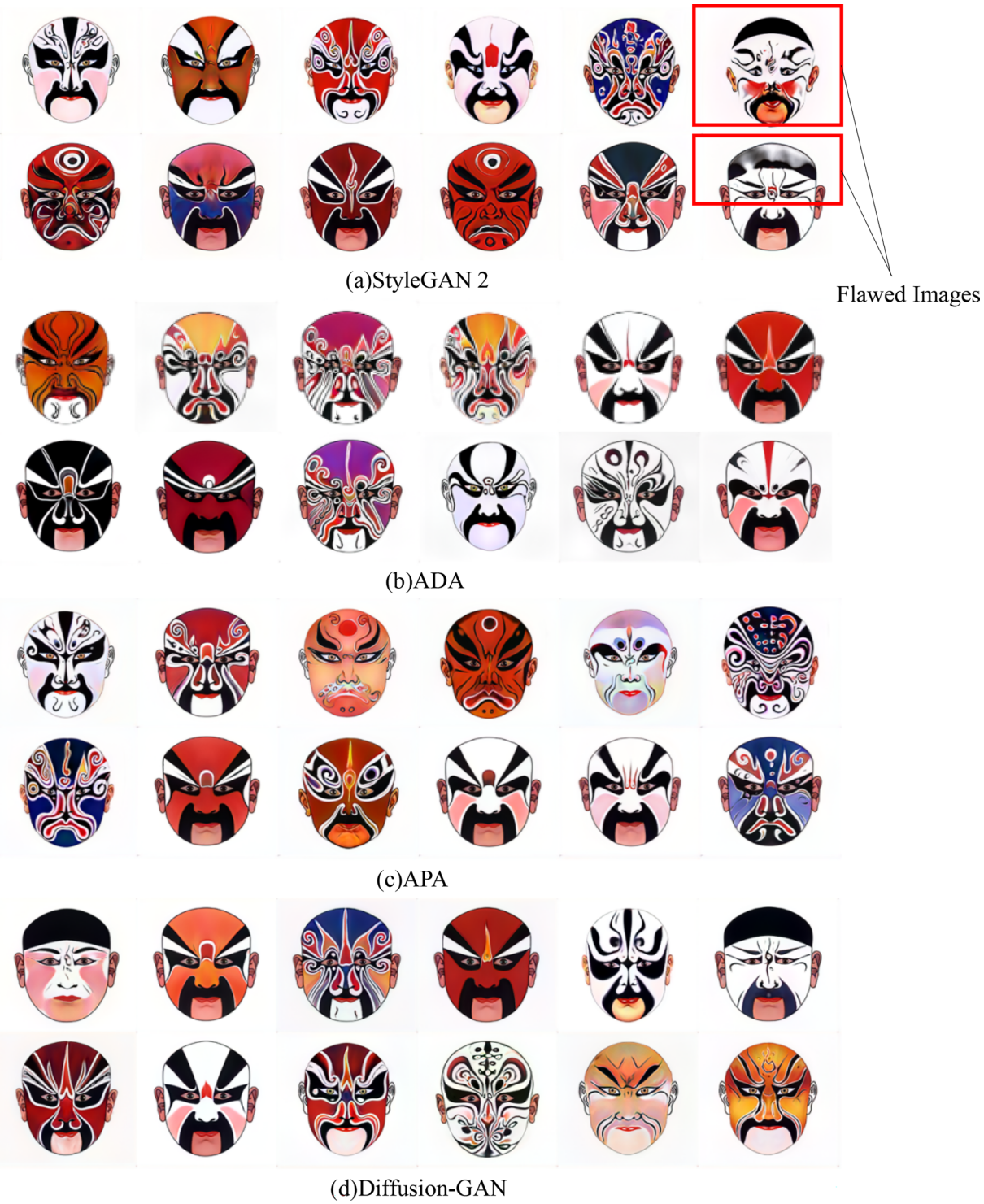

Fig. 3 Visual comparison of generated images. **a** Images generated by StyleGan2. (Defective images are marked with red rectangular boxes.) **b** Images generated by StyleGan2-Ada. **c** Images generated by StyleGAN2-APA. **d** images generated by Diffusion-GAN

Fig. 4 Peking Opera Facial Masks Images Generated by CoCosNET. **a** Line Drawings **b** Style References **c** Generated Images

Fig. 5 Visual comparison of images generated by both CoCosNET and SPADE models. (Red box marks the color blending area)

Table 1 Comparison of the quality of images generated by image generation models

Image generation models	FID	KID ($\times 10^{-3}$)	SSIM	MS-SSIM
StyleGAN2	21.63	14.94	**0.4981**	**0.2965**
ADA	**15.09**	**6.97**	0.4887	0.2801
APA	18.87	10.81	0.4902	0.2853
Diffusion-GAN	22.58	11.65	0.4963	0.2959

The best scores are highlighted

and technical support for the inheritance and preservation of Peking opera facial masks.

Quality assessment of generated images

In this paper, four image quality assessment metrics, Frechet Inception Distance (FID), Kernel Inception Distance (KID), Structural Similarity (SSIM), and Multi-Scale-Structural Similarity (MS-SSIM), are used to compare the quality of the Peking opera facial masks images generated by several methods applied in this paper [26–28]. FID represents the distance between the distribution of the generated images and the distribution of the real images. KID calculates the square of the maximum average difference between the feature representation of the real images and the feature representation of the generated images. Lower values of FID and KID represent better quality of the generated images, and as the values increase, the generated images become more distorted. SSIM measures the structural similarity between two images. MS-SSIM is a multi-scale based on the SSIM index. SSIM values and MS-SSIM values range from 0 to 1, and as the value increases, the more similar the images are to each other.

As can be seen from Table 1, the FID score and KID score of StyleGAN2-ADA are the lowest values among the four models. This proves that the network with the ADA module achieves the best generation effect. However, the experimental results of APA are unexpected. In theory, expanding the training set with a constant stream of generated dummy samples is a way to avoid model overfitting, stabilize training and improve the quality of synthetic images. The reason why the improvement effect is not so amazing may be that the addition of false images interferes with the normal training of the discriminator. Diffusion-GAN becomes worse with the same number of trainings. We speculate the main reason is that discriminator learning not only distinguishes between true and false data but also needs to distinguish diffusion-generated samples from diffusion-true observations. In addition, the injection of instance noise aggravates the learning task of model training, which prolongs the training time, although avoids catastrophic forgetting of the discriminator to some extent.

In addition, we note that the StyleGAN2 model has the highest SSIM value and MS-SSIM value among the four models. This indicates that the images generated by the StyleGAN2 model are most similar to the real images, while the images generated by the network with the addition of the data enhancement module are less similar to the real images. This is because the real data distribution in GAN training will be disturbed by adding a data enhancement module.

Table 2 shows a comparison of the image quality between our adopted image translation model and another advanced image translation method, SPADE. We can see that the method used in this paper is significantly better than SPADE.

Comparing Table 1 and Table 2, we can see that the image translation method is obviously superior to the image generation method in this experiment. Both the FID score and KID score of CoCosNET are lower than StyleGAN2-ADA which is the most effective of the image generation methods. We speculate that it is because the image translation uses the idea of coloring the shape structure of the Peking opera facial masks, which makes the generated images more detailed and accurate than the fake images generated directly by the image generation model. The SSIM and MS-SSIM values of the CoCosNET model are higher than those of image generation models because the structures of images generated by CoCosNET are all from real images.

Table 2 Comparison of the quality of images generated by image translation models

Image translation models	FID	KID (*10^{-3})	SSIM	MS-SSIM
SPADE	58.03	44.13	0.3423	0.1461
CoCosNET	**8.988**	**2.89**	**0.850**	**0.896**

The best scores are highlighted

Diversity evaluation of generated images

In this paper, Learned Perceptual Image Patch Similarity (LPIPS) is used to measure the average feature distance between the generated samples. LPIPS represents the perceptual similarity between image blocks and is used to measure the difference between two images [29]. The lower the LPIPS score, the less the difference between the images. In other words, the higher the LPIPS score, the better the diversity among the generated image samples.

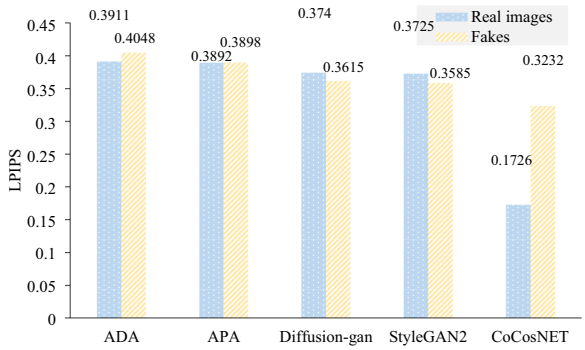

Fig. 6 LPIPS values of the generated images from different models. The blue color represents the LPIPS values between generated images and real images, and the yellow color represents the LPIPS values between generated image samples

As can be seen from Fig. 6, the diversity of image samples which are generated by the image generation model is higher than the diversity of image samples generated by the image translation model. The similarity between the images generated by the image translation model and the real images is higher. This is because the image translation model generates the image structures from the real images. Therefore, the generated images are more similar to the real images, but the price is that the diversity of the generated images will be reduced. The image generation model learns the distribution of real images and generates new images, so the diversity of generated samples is higher.

On the other hand, we can also see that the diversity of image samples generated by the StyleGAN2 network with a data enhancement module is higher than that without data enhancement. This indicates that data enhancement is an effective method to improve the diversity of the generated Peking opera facial mask images.

Classification task

In order to further evaluate whether the Peking opera facial mask generation task is up to standard, and compare the quality of Peking opera facial mask images generated by image generation and image translation, this paper uses an image classification task to verify. The main color of Peking opera facial masks represents the character and significance of the characters. In this paper, the Peking opera facial masks in the original dataset are classified into seven categories according to different primary colors. The classification criteria are shown in Fig. 7. After the manual screening, 500 Peking opera facial masks of each type are extracted from the original dataset as the classification training set and 20 Peking opera facial masks of each type are obtained from the remaining dataset and the images generated by the model studied in this paper as the classification test set.

Considering that data sets are classified by color, the classification task is relatively simple, so the required network layers need not be too deep. In this paper, a famous convolutional neural network, ResNet18, is adopted as the feature classifier. The transfer learning technology is used to modify the output category of the classification layer. The pre-training network is used to initialize the network. The last fully connected layer is replaced by a new layer with random weight, and only this layer is trained. In this paper, the batch_size is set to 8 and the num_epochs is set to 50. Table 3 shows the performance comparison between the research method in this paper and the original data set after training through feature extraction on the classification model.

Taking the classification accuracy of the original datasets as the standard, we find that the classification accuracy of the images generated by CoCosNET is closest to the classification accuracy of the dataset (real images) and is the highest among the research methods in this paper.

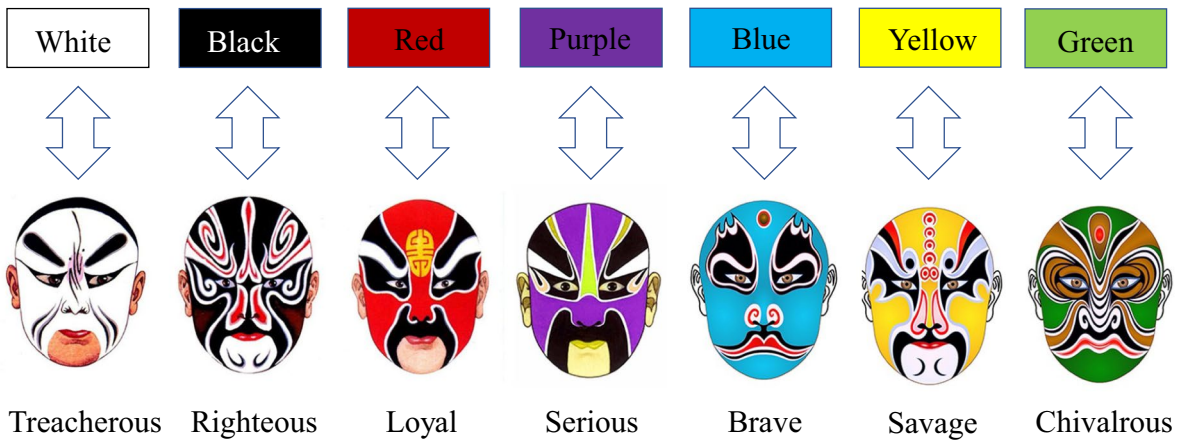

Fig. 7 Criteria for classifying Peking opera facial masks

Table 3 Comparison of classification accuracy of images generated by different models

Model	Real datasets	ADA Fakes	APA Fakes	Diffusion-GAN Fakes	StyleGAN2 Fakes	CoCosNET Fakes
Acc	94.2%	90.5%	89.7%	79.8%	81.4%	**92.9%**

(The best scores are highlighted)

This illustrates that the images generated by CoCosNET are closest to real images and can be used in place of real images for Peking opera facial mask character classification applications. At the same time, it also proves that for Peking Opera face images, the quality of the images generated by the image translation technique which is based on the instance images is better than that generated by the image generation technique.

On the other hand, the classification accuracy of the images generated by StyleGAN2 adding the data enhancement modules ADA and APA is much higher than that only generated by StyleGAN2. It demonstrates that applying the data enhancement modules to the StyleGAN2 network can reliably stabilize the training and effectively improve the quality of image generation.

Conclusion

The construction of the cultural gene pool of Peking opera facial masks and the automatic generation of Peking opera facial masks are technological innovations in traditional arts. In this paper, we explore the intelligent generation method of Peking opera facial masks from two directions: image generation and image translation. The generated images have achieved good results both in qualitative analysis and quantitative analysis. This has important practical implications for promoting both the dissemination and reuse of Peking opera cultural resources.

In terms of image generation, this paper explores the best data processing in the generative model by applying different data enhancement methods on the StyleGAN2 network. Our experiments show that data enhancement remains the simplest and most effective way to avoid discriminator over-fitting and to improve the quality of generated images. This will lay a good foundation for Peking opera facial mask generation, including improving network structure and training strategies. In terms of image translation, the CoCosNET achieves a good generation of Peking opera facial masks, but the problem of color mixing in some regions occurs in the generation of certain complex spectral patterns. The quality of the generated images can be further improved if the dataset is finely chunked and calibrated.

We intend to focus our future work on building mask datasets of Peking opera facial mask segmentations and researching data enhancement. Combining these two aspects to improve the network structure, further improving the quality of Peking Opera facial masks, and providing reliable technical support for the development and protection of Peking Opera facial masks.

Abbreviations

StyleGAN	Style generative adversarial network
GAN	Generative adversarial network
AR	Augmented reality
VAE	Variational auto encoder
ResNet	Residual networks
3D	Three-dimensional
APA	Adaptive pseudo augmentation
ADA	Adaptive data augmentation
FID	Frechet inception distance
KID	Kernel inception distance
SSIM	Structural similarity index
MS-SSIM	Multi-scale-structural similarity index
LPIPS	Learned perceptual image patch similarity

Acknowledgements

None.

Author contributions

All the authors contributed to the current work. YM and XR devised the study plan and led the writing of the paper. SYH and XR arranged the data of experiment. YM and JC supervised the entire process and provided constructive advice. All authors read and approved the final manuscript.

Funding

This paper was supported in part by the Open Project of Key Laboratory of Audio and Video Repair and Evaluation, Ministry of Culture and Tourism (Grant No. 2021KFKT007), and the Fundamental Research Funds for the Central Universities.

Availability of data and materials

All data for analysis in this study are included within the paper.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 7 December 2022 Accepted: 13 January 2023
Published online: 30 January 2023

References

1. Tu H. The cultural connotation and symbolic meaning of Chinese opera mask color. 2016 3rd International Conference on Education, Language, Art and Inter-cultural Communication (ICELAIC). Atlantis Press. 2016. p. 466–468
2. Xu D, Nie Z, Zhou W. From Traditional Culture Education, the Application of Peking Opera Facial Elements in Poster Design Teaching. International

Conference on Education and Management (ICEM). Atlantis Press. 2018. p. 855–858
3. Santoso DJ, Angga WS, Silvano F, Anjaya HES, Maulana FI, Ramadhani M. Traditional mask augmented reality application. 2021 International Conference on Information Management and Technology (ICIMTech). IEEE. 2021. p. 595–598
4. Pratama D, Karya SV, Maulana FI, Ramadhani M, Permana F, Pangestu G. Introduction to mask Malangan with augmented reality technology. 2021 International Conference on Information Management and Technology (ICIMTech). IEEE. 2021. p. 364–368
5. Liu K, Gao Y, Zhang J, Zhu C. Study on digital protection and innovative design of Qin opera costumes. Herit Sci. 2022;10:127.
6. Yan M, Wang J, Shen Y, Lv C. A non-photorealistic rendering method based on Chinese ink and wash painting style for 3D mountain models. Herit Sci. 2022;10:186.
7. Ali H, Biswas MR, Mohsen F, Shah U, Alamgir A, Mousa O, Shah Z. Correction: the role of generative adversarial networks in brain MRI: a scoping review. Insights Imaging. 2022;13:125.
8. Yan M, Lou X, Chan CA, Wang Y, Jiang W. A semantic and emotion-based dual latent variable generation model for a dialogue system. CAAI Trans Intell Technol. 2023;2023:1–12. https://doi.org/10.1049/cit2.12153.
9. Tang H, Xu D, Sebe N, Wang Y, Corso JJ, Yan Y. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019. p. 2412–2421
10. Hu M, Guo J. Facial attribute-controlled sketch-to-image translation with generative adversarial networks. J Image Video Proc. 2020;2020:2.
11. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T. Training generative adversarial networks with limited data. Adv Neural Inf Process Syst. 2022;2022:33.
12. Lv C, Li Z, Shen Y, Li J, Zheng J. SeparaFill: Two generators connected mural image restoration based on generative adversarial network with skip connect. Herit Sci. 2022;10:135.
13. Zhang H, Sindagi V, Patel VM. Image de-raining using a conditional generative adversarial network. IEEE Trans Circuits Syst Video Technol. 2020;30:11.
14. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2020. p. 8107–8116
15. Ma S, Cao J, Li Z, Chen Z, Hu X. An improved algorithm for superresolution reconstruction of ancient murals with a generative adversarial network based on asymmetric pyramid modules. Herit Sci. 2022;10:58.
16. Jiang L, Dai B, Wu W, Loy CC. Deceive D: Adaptive pseudo augmentation for gan training with limited data. Adv Neural Inf Process Syst. 2021;2021:34.
17. Wang Z, Pavan FRM, Sayed AH, Decentralized gan training through diffusion learning. IEEE 32nd international workshop on machine learning for signal processing (MLSP). IEEE. 2022;2022:1–6.
18. Tang H, Liu H, Sebe N. Unified generative adversarial networks for controllable image-to-image translation. IEEE Trans Image Processing IEEE. 2020. https://doi.org/10.1109/TIP.2020.3021789.
19. Wang M, Yang GY, Li R, Liang RZ, Zhang SH, Hall PM, Hu SM. Example-guided style-consistent image synthesis from semantic labeling. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019. p. 1495–1504
20. Park T, Liu MY, Wang TC, Zhu JY. Semantic image synthesis with spatially-adaptive normalization. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019. p. 2332–2341
21. Zhang P, Zhang B, Chen D, Yuan L, Wen F. Cross-domain correspondence learning for exemplar-based image translation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2020. p. 5142–5152
22. Bora A, Jalal A, Price E, Dimakis AG. Compressed sensing using generative models. 2017 34th International Conference on Machine Learning. PMLR. 2017. p. 537–546
23. Yang Q, Yu HX, Wu A, Zheng WS. Patch-based discriminative feature learning for unsupervised person re-identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2019. p. 3628–3637
24. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2017. p. 2117–2125
25. Winnemöller H, Kyprianidis JE, Olsen SC. XDoG: An eXtended difference-of-Gaussians compendium including advanced image stylization. Comput Graph. 2012;36:6.
26. Chong MJ, Forsyth D. Effectively unbiased fid and inception score and where to find them. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. 2020. p. 6069–6078
27. Dimitriadis A, Trivizakis E, Papanikolaou N, Tsiknakis M, Marias K. Enhancing cancer differentiation with synthetic MRI examinations via generative models: a systematic review. Insights Imaging. 2022;13:188.
28. Dost S, Saud F, Shabbir M, Khan MG, Shahid M, Lovstrom B. Reduced reference image and video quality assessments: review of methods. J Image Video Proc. 2022;2022:1.
29. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. 2018. p. 586–595

Publisher's Note