The Visual Computer (2024) 40:5451–5463
https://doi.org/10.1007/s00371-023-03115-2

ORIGINAL ARTICLE

CA-GAN: the synthesis of Chinese art paintings using generative adversarial networks

Zihan Chen[1] · **Yi Zhang**[1]

Accepted: 13 September 2023 / Published online: 16 October 2023

Abstract
With the advent of generative adversarial networks (GAN), an astonishing advancement has been made in the generation of art painting in recent years. However, existing methods still suffer problems such as color confusion or blurred details. In addition, most of those works centered around the generation of western art painting, while less attention was paid to Chinese traditional arts. Moreover, the lack of traditional Chinese painting datasets is also one of the reasons for the delayed development. To solve the above problems, our research focuses on the synthesis of multi-style traditional Chinese paintings. Firstly, we collect and sort out more than 1000 traditional Chinese paintings, including line drawings, meticulous paintings, ink paintings. Secondly, we propose a Chinese art generative adversarial network (abbreviated as CA-GAN) to decouple the latent vector based on attention mechanism. CA-GAN maps an image to content space and attribute space and fuses them to generate high-quality traditional Chinese art paintings. Meanwhile, a content discriminator is presented to check the consistency of mapping process based on cross-cycle consistency constraint. To make the generated images more artistic, MS-SSIM loss and Charbonnier loss functions are adopted to improve the performance of our model. Experiments have been conducted to verify the effectiveness and the generalization ability of our model. Compared with other state-of-the-art methods, the Chinese art paintings generated by CA-GAN are more vivid and realistic, and the resolutions of them are increased to 280 × 280.

Keywords Image synthesis · Generative adversarial networks (GAN)

1 Introduction

The traditional Chinese painting has drawn more and more attention due to its high ornamental value and artistic nature [1, 2]. It has various types and styles, including line drawings, meticulous paintings and ink paintings, etc. Meanwhile, the painting style is also diverse, ranging from vivid realistic figure paintings to abstract painting of flowers and birds embellished by colors. Out of the artistic pursuit of these traditional Chinese paintings with rich artistic connotation and property, scholars attempted to create effective models for machines to learn to draw those paintings by themselves [3]. However, each painting has their unique style, especially the characteristics of Chinese style and complex structures in the paintings are normally difficult to capture and imitate. Fortunately, the image-to-image translation technique has been widely studied with outstanding achievement, which facilitates the synthesis of traditional Chinese paintings.

In early years, the image translation tasks were mainly accomplished by style transfer. With the powerful feature extraction abilities of convolution neural networks (CNNs), the color features and drawing details were brought into the real photographs through neural style transfer [4]. Later, new ideas for end-to-end artistic image generation tasks were developed on the basis of Variational Auto-Encoder (VAE) [5], in which the mapping from the source dataset to the target dataset was learned to transform the unknown samples to real images. With the proposal of generative adversarial networks [6] (GAN), researchers were able to obtain high-quality synthesized images by optimizing the model structure and loss functions of GAN based on the adversarial training theory between the generator and discriminator. Since then, many works have been published studying the synthesis of traditional Chinese painting using GAN. Lin et al. [7] trained their model on multi-scale images and transformed simple sketch

✉ Yi Zhang
yi.zhang@scu.edu.cn

Zihan Chen
chentzuhan@stu.scu.edu.cn

[1] Department of Computer Science, Sichuan University, Chengdu, China

drawings into traditional Chinese paintings. ChipGAN [8] transformed real photographs into artistic paintings of Chinese ink style with superior visual quality and high stylization perceptual scores. Yu et al. [9] put forward a new framework for the synthesis of image-based Chinese landscape painting, which were much closer to manual work than previous methods.

Although many explorations have been made in the field of artistic image generation with proven achievements, the generated images were usually of low quality with blurred details. In addition, the standardized datasets about Chinese traditional painting are very scarce. These problems have further hindered the development of research on the generation of Chinese traditional painting. To address the above problems, we propose CA-GAN to generate vivid and artistic traditional Chinese paintings, with special emphasis flower-bird paintings. CA-GAN is able to convert the original painting style between line drawing, meticulous painting and ink painting and specify the style of generated images. To this end, CA-GAN separates the content space and attribute space of an image and encodes the paintings into the content space and attribute space, respectively, for better extraction of advanced features. Besides, we introduce attention mechanism into our model to create images with finer features, so as to better restore the details of the flowers and birds. We also adopt a U-Net [10]-like symmetrical cascade structure as the feature generator, which consists of multiple convolution and residual blocks. And a traditional convolution structure is utilized as the discriminator. To make the style of generated images more diverse, we utilize MS-SSIM loss to reinforce the cross-loop consistent constraints and adopt Charbonnier loss to alleviate the problem of insufficient model diversity caused by traditional L2 loss for image reconstruction. Figure 1 shows the traditional Chinese paintings created by CA-GAN. Experimental results (in Sect. 4) demonstrate the effectiveness of our model and its applicability for related style transfer tasks.

In a nutshell, our main contributions in this paper are summarized as follows:

1. We propose CA-GAN, a new network with separated content and attribute spaces, which creates decoupled representation of latent space via attention mechanism.
2. We add random Gaussian noise to CA-GAN as a feature vector to make the synthesized traditional Chinese painting more artistic and diverse.
3. We increase the resolution of the synthesized images from 256×256 to 280×280.
4. We employ Charbonnier loss and MS-SSIM loss for image reconstruction and reap the benefit of cross-cycle consistency constraint to ensure the quality of the synthesized images.
5. We collect and sort out datasets of Chinese line drawing, meticulous painting and ink painting to facilitate further future research, which can be downloaded via https://pan.baidu.com/s/1-vMF4eXMejboG9DP2ghuLg?pwd=t4ig. The unzip password for this file is "ChinesePaintings268."

The rest of the paper is organized as follows: related works are discussed in Sect. 2. The architecture of our network is described in Sect. 3. The experimental results with ablation studies are shown in Sect. 4 with thorough analysis. A final conclusion is drawn in Sect. 5.

2 Related works

2.1 Generative adversarial networks

As a prevalent method, generative adversarial network (GAN) has been widely applied to artistic style transfer tasks [4] (e.g. turning photographs into paintings, or creating paintings). GAN consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$,

Fig. 1 The images created by CA-GAN. The first row shows the meticulous paintings; the second row shows the ink style paintings

where $\mathcal{G}$ is used to generate real samples from Gaussian random noise, trying to keep the generated sample as consistent with the actual sample as possible, so that $\mathcal{D}$ cannot distinguish. On the other hand, the discriminator $\mathcal{D}$ aims to identify whether the sample is real or is generated by $\mathcal{G}$. The generator and the discriminator compete with each other until the discriminator finally cannot distinguish the authenticity of the samples. The classical loss function of GAN is written as 1:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\mathcal{D}, \mathcal{G}) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log \mathcal{D}(\boldsymbol{x})] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{z})))] \tag{1}$$

After the original GAN, many improved versions also came about to boost its performance and broaden the scope of its application scenarios. In the meantime, researchers are committed to optimizing the structure of generator and discriminator, loss functions and improving the training techniques, etc.

Network architecture: Larsen et al. [11] proposed a combination of Variational Autoencoder (VAE) [5] and GAN (dubbed VAE-GAN), wherein the generator acts as part of the decoder of the VAE to improve the overall performance through reconstruction loss function. Bao et al. [12] elaborated CVAE-GAN, which modeled an image as a composition of label and latent attributes in a probabilistic sequence.

The endeavor to integrate attention mechanism and GAN is also an innovative work to improve the performance of GAN [13–17]. Attention-GAN [13] pointed out that generators could not perceive the most discriminative parts between source and target domain and thereby created an attention mask to fuse the generator output to obtain the high-quality images. AdaLIN [15] advocated a new attention module with normalization function, which weighted the feature maps with fully connected weights by auxiliary classifiers of the source and target domains. It also combined the attention module to control the degree of changes in style and content in an adaptive way. TransGAN [18] was a pioneering work in building a GAN without convolution blocks, which contains a transformer-based [19, 20] memory-friendly generator that increases feature resolution in a progressive way and a discriminator to capture semantics and low-level textures.

Loss functions: Since GAN has been proposed, the gradient instability and mode collapse problems have always puzzled researchers. It is well known that during training process, the better the discriminator, the worse the vanishing gradient problem of generator will be (i.e., it is hard for the generator to learn the distribution of the original samples). In addition, the unjustified distance measure between 2 distributions often leads to the insufficient model diversity dilemma.

Considering this, Arjovsky et al. [21] mathematically analyzed the reason for the instability of GAN training and suggested the use of approximate Wasserstein distance to measure 2 distributions. WGAN [21] made theoretical analysis toward fully understanding the training dynamics of GAN to improve stability. But their quality of generated samples was low, and the model was difficult to converge. Gulrajani et al. [22] declared that WGAN's weight clipping strategy satisfied the Lipschitz constraint of the discriminator forcibly, leading to an uncontrollable training process. Therefore, a truncation scheme with gradient penalty was adopted to stabilize the training of WGAN. These works relieved the unstable training problem to some extent and provided corresponding metrics to measure the training process.

Nowozin et al. [23] claimed that any divergence (collectively referred to as f-Divergence) can be used in GAN models to measure the distance between distributions. LSGAN [24] minimized the Pearson χ^2 divergence by using a least squares loss for the generator. The effectiveness of LSGAN had been proven by a series of experiments, and the quality of generated samples and stability of training had been significantly improved compared with WP-GAN [22].

2.2 Image-to-image translation

The image-to-image translation technique has been widely applied in a variety of applications. Typical examples include day-to-night photograph conversion, artistic style transfer and line coloring, etc. The essence of image-to-image translation is to let the machine learn the mapping rules between given image pairs, so that it could fulfill the task of mapping an image from a source domain to a target domain. Besides, both the source and the generated images are required to share the same distribution. As a sub-branch of image-to-image translation, style transfer has also made tremendous progress. Gatys et al. [4] rendered the semantic content of an image by using the image representations derived from convolution neural networks in an explicit way to generate artistic photographs. However, this method is computationally heavy, and some of the generated details are blur.

GAN has been the de-facto standard for image-to-image translation with appealing results. pix2pix [25] not only learned the mapping from the input to output images, but also learned a loss function to train this mapping. However, it is often difficult to gather paired images in practice, and the generalization ability is quite limited with weak diversity. CycleGAN [26] realized the translation from source domain to target domain without the paired samples, but the learning of style information turned out to be insufficient. UNIT [27] and MUNIT [28] addressed the above-mentioned problems in CycleGAN, which mapped the images pairs into a shared latent space.

3 Method

In general, traditional Chinese paintings reflect the culture and customs of ancient China, in which the specific artistic details are inseparable from its styles, expressing the painter's ideological and emotional conception. The meticulous paintings emphasize specific expression, while ink paintings emphasize abstraction. In terms of the synthesized images, the word "Artistry" means higher resolutions, rich details and multiple available styles to express the artistic nature.

Our main goal of this paper is to learn the mapping between different styles of traditional Chinese flowers & birds paintings. Although previous works generated vivid Chinese paintings, they did not express the artistry quite well. We believe the reason is that those models failed to decouple the artistic features and the content features of Chinese paintings. In this light, we propose CA-GAN, which encodes [27–30] an image into 2 vectors through encoding, where 1 vector represents its content, and another 1 characterizes its style. Meanwhile, CA-GAN also encompasses attention mechanism to let the network focus on birds and flowers in the painting to facilitate translation work.

3.1 CA-GAN

The network architecture of CA-GAN is illustrated in Fig. 2, which is comprised of multiple encoders, generators & discriminators and attention modules. We use X to represent line drawing, and Y for meticulous painting, and x and y denote the samples of each of them, respectively. In order to let the generator to fully learn the styles of the images in 2 domains so as to better integrate their features, we implement image translation via 2 stages. And the entire model is trained end-to-end.

We also introduce attention mechanism [16] into our network, so as to make the generated images focus more on the foreground elements (i.e., birds and flowers), while less on background scenes. To be specific, the samples x and y are firstly sent to the attention module A_X to enhance the representation of the foreground objects as $x_a = A_X(x)$, and the result is then sent to the encoder.

Our attention block is shown in Fig. 2, which is constructed as an encoder–decoder structure along with multiple convolution and residual blocks. It is worth noting that the x in Fig. 2 (b) and Fig. 2 (c) does not refer specifically to the image belonging to the X domain, but to all the images entered into the attention module. The size of the attention map x_a (output by the attention module) is equal to that of the original sample x, which are all normalized into [0, 1].

Step 1: Image synthesis

As shown in Fig. 2, the transformation of domain X to domain Y is implemented as follows: Firstly, the sample x is sent to the attention block to yield the attention map x_a. Next, x_a and x are fused to obtain the enhanced foreground map x_f, as shown in equation 2:

$$x_f = x \odot x_a \tag{2}$$

Then x_f is fed to the attribute encoder E_X^a and the content encoder E_X^c in the image domain X individually to produce encoded attribute z_{xf}^a and encoded content z_{xf}^c, respectively. A similar process is carried out for sample y in image domain Y at the same time to create z_{yf}^c and z_{yf}^a. That is to say, we map an image to content space and attribute space, so that we can represent one image with both content and attribute features.

Then we swap and fuse the above-mentioned hidden vectors as follows: Firstly, we concatenate z_{xf}^a and z_{yf}^c (denoted by "circled C" in Fig. 2) and send the concatenated vector to the generator G_X to generate the image x'_f in domain X (And y'_f is generated similarly by G_Y). As described earlier, the images generated by attention block only contain foreground objects (refers to [16]). To attain a complete image, we need to fuse the background elements through background synthesis:

$$\begin{aligned} x' &= y_a \odot x'_f + (1 - y_a) \odot x \\ y' &= x_a \odot y'_f + (1 - x_a) \odot y \end{aligned} \tag{3}$$

The domain discriminators D_X and D_Y are used to determine if the output image belongs to domain X or Y, respectively.

In addition, we also advise a content discriminator D_c to measure the consistency of the encoded content features in the 2 domains. Because the feature of contents (birds and flowers) should be identical, even if the style of the images is different.

After the above steps, sample x is mapped to attribute space A_X and content space C of domain X through encoder E_X^a and E_X^c, respectively. And sample y is mapped to attribute space A_Y and the same content space C accordingly.

Step 2: Verification of synthesized image

We impose the cross-cycle consistency constraint to check the correctness of the mutual mapping between images in domain X and Y.

In step 1, x and y are encoded into common content space and unique attribute space, respectively, and their attribute spaces are also swapped and decoded to yield new images x' and y'. In this step, x' and y' are treated as inputs, which are processed the same way as in step 1, and are decoupled by the encoder to extract the latent vectors in both attribute space and content space. Again, a similar swapping process (as described in step 1) is carried out to obtain images of domain X and Y, written as x'' and y'', respectively.

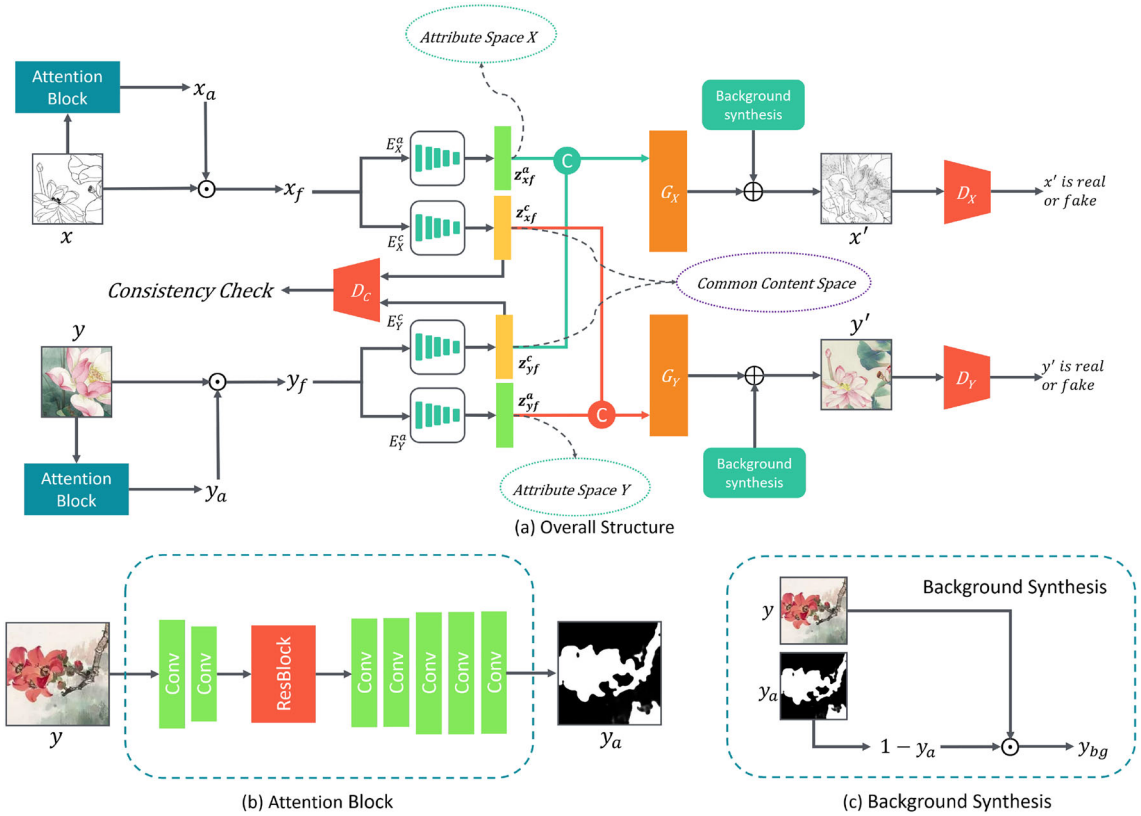

Fig. 2 The overall architecture of CA-GAN. **a** shows the mapping process by the attention module, where the images of two domains are mapped to attribute and content spaces, respectively, through the exchange of their hidden feature vectors. **b** shows the creation of attention image, while **c** illustrates the synthesis of background image

According to cross-cycle consistency theory, the reconstructed input image x and output image y should be identical to the original images after 2 swapping of encoding space, i.e., $x = x''$ and $y = y''$, and we use a loss function L_{cc} to validate this result.

Traditionally, for example, an input image x (belongs to X domain) with the attribute vector extracted by the attribute encoder of domain Y from style-image y (belongs to Y domain) could not create diversity style. To produce diverse and visually appealing artistic images, we introduce Gaussian noise in the training process (as a special type of multi-style vector). To be specific, we generate a random Gaussian distributed noise z in the latent space, which will be fed to the generator as the attribute representation of the image (to replace the attribute representation extract from attribute encoder of domain Y). Then our decoder re-builds the noise, which means that if an image is generated by the random vector as its attribute vector, we then require that after this image is mapped back by the attribute encoder of Y domain into the input noise, it still equals to the noise that we input. Most existing methods focused on the generation of fixed style images; they translated the style image into the corresponding attribute vector without noise (even if they infuse noise into the input, the generator would simply ignore).

After training, we can testify the completion of the image conversion process. The input image x is firstly sent to an attention block to strengthen the foreground elements and is then fed to the content encoder E_X^c of domain X to obtain the content feature z_{xf}^c of image x. z_{xf}^c is sent to generator G_Y together with the noise (or the attribute feature vector of the real sample of domain Y, i.e., z_{yf}^a)

Considering we have aligned the attribute feature vectors of the real samples in domain Y with Gaussian distribution during training, thus, we can either generate an image with random Gaussian distributed noise as attribute vector or specify the specific style with the attribute vector from the real samples of domain Y.

3.2 Loss function

We propose 3 loss functions in our model, namely adversarial loss, cross-cycle consistency loss and reconstruction loss.

3.2.1 Adversarial loss

Our adversarial loss includes content adversarial loss and attribute adversarial loss. For different styles of Chinese paintings, the representation of the main elements in the paintings could be encoded into a common hidden vector space. Thus, we utilize the discriminator D_c to distinguish the implicit vectors that are encoded by 2 images in different domains, so that they are forced to be mapped to the same content space. For the 2 content encoders and content discriminators, the content adversarial loss is expressed as:

$$\begin{aligned} L_{ac} = \mathbb{E}_x \left[\frac{1}{2} \log D^c \left(E_X^c(x) \right) + \frac{1}{2} \log \left(1 - D^c \left(E_X^c(x) \right) \right) \right] \\ + \mathbb{E}_y \left[\frac{1}{2} \log D^c \left(E_Y^c(y) \right) + \frac{1}{2} \log \left(1 - D^c \left(E_Y^c(y) \right) \right) \right] \end{aligned} \tag{4}$$

The discriminators D_X and D_Y are applied (with a conventional discriminator loss L_{ad}) to determine if the images generated by G_X and G_Y belong to their target domains. Here, we employ least square loss to reduce instability in training.

3.2.2 Cross-cycle consistency loss

For cross-cycle consistency loss, since we have represented the image in both shared content space and different attribute spaces, image x' and y' are obtained through swapping their attribute and content representations. We feed x' and y' into the encoder again to encode their contents and attributes separately and swap to generate new corresponding image pair x'' and y''. Apparently, after the verification process, x'' and y'' are transformed back to the original sample x and y (after the restoration process of attention block). It is worth noting that cross-cycle consistency loss is not a direct cycle consistency reconstruction of the image (like $X \rightarrow Y \rightarrow X$ in CycleGAN [26]), but a cross-cycle reconstruction of the disentangled representation of the content and attribute. Assume the cross-cycle reconstruction is denoted as L_{cc}, which is used to ensure the consistency of the cross-cycle reconstruction of the content and attribute of the disentangled representation.

Here, we use L1+MS-SSIM [31] to check image consistency. Formally,

$$\begin{aligned} L_{cc} = \mathbb{E}_{x,y} \left[\left\| G_X \left(E_Y^c(y'), E_{\mathcal{X}}^a(x') \right) - x \right\|_1 \right. \\ \left. + \left\| G_Y \left(E_X^c(x'), E_Y^a(y') \right) - y \right\|_1 \right] \\ + F_{\text{ms-ssim}}(x, x') \end{aligned} \tag{5}$$

Here x' and y' are generated images after the swapping of attribute space and content space, respectively.

However, SSIM [32] has long been proposed to compare the brightness, structure and contrast of pictures, which alleviates the lack of diversity of the model. However, since it emphasizes the global features of an image, it often overlooks the details. Under such circumstances, we apply MS-SSIM [31] to remedy the above deficiency, by paying more attention to the local features of an image:

$$\begin{aligned} \text{MS} - \text{SSIM}(x, y) = \\ [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \end{aligned} \tag{6}$$

3.2.3 Reconstruction loss

In step 1 of Sect. 3.1, we performed cross-domain transformation operation by swapping attribute encoding and content encoding for 2 domains. If we simply fuse them and send the result to the generator, we get the reconstructed image as: $\hat{x} = G_X(z_{x_f}^a, z_{x_f}^c)$, $\hat{y} = G_Y(z_{y_f}^a, z_{y_f}^c)$. Therefore, to recover the original image, we need to integrate the attribute features and content features to reach the desired outcome: $\hat{x} = x$, $\hat{y} = y$. Considering the traditional L2 loss for pixel-level comparisons lack diversity in the generated images [33], we thereby use the reconstruction loss L_{rec} to further improve the quality of the generated images, which is based on Charbonnier Loss:

$$L_{\text{rec}} = (\hat{x}, x) = \frac{1}{hwc} \sum \sqrt{(\hat{x} - x)^2 + \epsilon^2} \tag{7}$$

Then, our final objective loss is calculated as:

$$L = L_{ac} + L_{ad} + \lambda_1 L_{cc} + \lambda_2 L_{\text{rec}} \tag{8}$$

here λ_1 and λ_2 are hyperparameters, which are used to balance the weights of L_{cc} and L_{rec}.

4 Experiments

In this section, we firstly describe the experimental details, including pre-processing and datasets involved. Then, we compare our result with other prevalent methods, including CycleGAN [26], UNIT [27], AGGAN [16] and FlowerGAN [34], wherein we use FlowerGAN as our baseline. Ablation study is conducted to verify the effectiveness of core components in our network along with the proposed loss functions. Finally, a human test is carried out to assess the quality of the generated images.

4.1 Implementation details

We implement CA-GAN using PyTorch 1.8.4 with Adam Optimizer [35] for gradient descent optimization. During training, all images are unpaired, we train our model to accomplish style transfer from line drawing to meticulous

Fig. 3 A qualitative comparison among several methods

drawing or to ink painting. For dataset, we create a new dataset of Chinese art painting by collecting and sorting out 1000 images for each category. And we crop and transform these images for data augmentation purpose. Then we normalize the RGB scale values of all images to $[-1, 1]$ and resize them into 280×280 (resolution). It is worth noting that the most commonly used input size for other methods is 256×256 (then the size of the corresponding output is also 256×256). For GAN structure, the input size cannot be infinitely large. Therefore, a very large input will cause unstable training of the model, which ultimately lead to mode collapse. Considering this, we use adaptive average pooling in our architecture so as to enable our model to take inputs of any scales. Specifically, we employ Spectral Norm Layer, Multi-scale Discriminator and Mode Seeking Regularization, etc., to stabilize the model training. Through experiment, we find that 280×280 is the highest resolution for stable training without negative effects (e.g., vanishing gradient or mode collapse).

4.2 Quality evaluation

A quality assessment of traditional Chinese paintings generated by CA-GAN is performed from 2 aspects. The assessment will demonstrate the effectiveness of our model in visual perception of artistry. Comparisons are made between our results and other popular methods.

4.2.1 Qualitative evaluation

A qualitative comparison is made and shown in Fig. 3. The first column is the input line drawing images of flowers. The first row on the right-hand side displays the results done by existing methods. The second row shows the provided stylized image and the corresponding results by our model (based on the styles). Firstly, from human visual perspective, our generated images are more vivid and artistic. Other models either suffer from blurred details or fail to express the classical characteristics of traditional Chinese paintings. Although FlowerGAN [34] better depicts artistic features in the generated images, the details of the flowers are not rich enough. Secondly, other models can only generate images of fixed styles, but we can create various styles according to different style clues. Thirdly, the resolution of our generate images (280×280) is higher than other methods (256×256). The increased resolution does not lead to model collapse or lack of delicacy. Instead, it boosts the visual effect. Finally, CA-GAN could generate multiple styles no matter for meticulous painting or ink painting, while other modes could only generate one style.

4.2.2 Quantitative evaluation

In quantitative experiment, we use inception score (IS) [36] and Frechet inception distance (FID) [37] as the metrics to evaluate the quality of the generated images. Inception score (IS) calculates the KL divergence between the conditional and marginal label distributions over the generated images, which describes the quality and diversity of the generated images. The score is produced by inception V3 network, which is developed by Google, and can recognize more than 1000 types of images. If a generated image has good quality and is clear enough, it should be easily recognized by inception V3. Therefore, we measure the quality and diversity of the generated images based on IS (the higher the quality, the higher the IS value). It is worth noting that the output of GAN networks has slight disturbance due to dropout. Thus, each pixel in the generated image could not be exactly the same. Nonetheless, such subtle difference would not affect the human visual perception. We use positive and negative signs ($\pm$) to represent this disturbance. FID score uses the distance of the image feature vector distribution to evaluate the quality of the image, and the lower the value, the higher the quality. The IS and FID scores of CA-GAN and other methods are shown in Table 1, in which CA-GAN achieves the best scores for both metrics.

Table 1 IS and FID metrics of state-of-the-art methods and CA-GAN

Method	IS ↑	FID ↓
CycleGAN [26]	3.175 ± 0.342	1.384
UNIT [27]	3.306 ± 0.648	1.984
AGGAN [16]	3.756 ± 1.285	1.285
FlowerGAN [34]	3.355 ± 0.312	1.272
CA-GAN (Ours)	3.506 ± 0.241	1.028

4.3 Ablation study

In this section, we conduct 2 ablation experiments. Firstly, we verify the effectiveness of the reconstruction loss and MS-SSIM loss, respectively. Secondly, we validate the impact of our proposed attention module.

The first column in Fig. 4 is the line drawing images. The second column shows the stylized meticulous drawings. The 3rd column shows the results by our model that does not contain the reconstruction loss. The 4th column is the results by our model that does not have MS-SSIM. The rightmost column is the results by our proposed model with MS-SSIM and reconstruction losses. Apparently, without MS-SSIM and reconstruction losses, the generated images are much inferior in terms of brightness, details, color saturation and artistry. The reason is that the model without MS-SSIM and reconstruction losses lacks perception of the artistry of traditional Chinese paintings and focuses less on the stylistic features of the input images. By contrast, our proposed model synthesizes images based on more accurate artistic representation and more realistic colors.

Compared with the 5th column and the rightmost column, the results without the attention module look more rigid than ours. In particular, the generated image in the 3rd row and the 5th column suffers color distortion (there should not be green and blue colors) due to mode collapse. The reason for this phenomenon is that our attention module captures the artistic details of the original stylized images, while those without the attention module lack such vividness.

4.4 The function of Gaussian noise

As explained in step 2 of Sect. 3.1, we introduce Gaussian distributed noise in the training stage as a special type of multi-style vector. As shown in Fig. 5, the provided input and style images are shown on the left. The generated images with/without Gaussian noise are shown on the right. Apparently, without the addition of Gaussian noise in the training process, the model can only generate images of fixed style, in

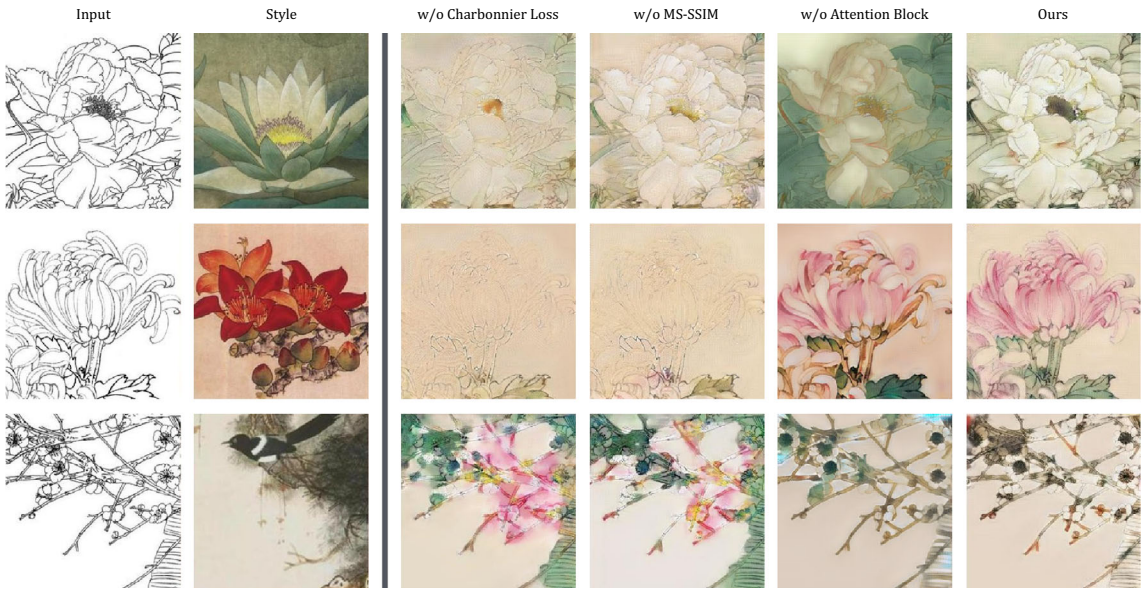

Fig. 4 Results of removing components in the loss function of our method

which the model will collapse if the noise is forcibly infused. By contrast, our proposed model (with Gaussian noise) is able to generate multi-style images based on different noises.

4.5 Attribute transition

In previous steps, we encode the images of 2 domains into a common attribute space and align the space with the prior Gaussian distribution. We can also perform image style transfer from random vectors sampled from the Gaussian distribution as attribute and linearly interpolate between the 2 attribute vectors, so as to generate images with new styles between attribute 1s and 2.

As shown in Fig. 6, the left column shows 2 line drawing images. The second and the last column show 2 meticulous paintings with attributes 1 and 2, respectively. The intermediate columns (from the 3rd to the 8th column) are the interpolation results of the 2 attributes. Since the encoding of the attribute space is continuous, the 2 different styles of attributes are transformed smoothly (from the left to the right). This results prove that our model is not only capable of encoding the styles of the image domain to a continuous attribute space, but could also create specific styles that do not exist in the target sample set based on randomly sampled vectors. This result also demonstrates the generalization ability of our model.

4.6 Human testing

To test and compare the visual perception of the images that are generated by our method and other methods, we randomly invited 100 people to do an image authenticity test. In this test, we compare 5 methods (CycleGAN, UNIT, AGGAN, FlowerGAN and ours), each of which generates 3 images (totally 15 images, and they are all fake and are mixed together, as shown in Fig. 7a). The participants need to judge whether the images are real or fake. The results are shown in Fig. 7b, the gray parts of the bar charts reflect the portion that the participants successfully distinguish that the images are fake, while other colors illustrate the proportions (of different methods) that the participants are fooled by the generated images (they mistakenly think they are real photographs). Apparently, the higher the color portions, the better the performance of the models. For example, the participants think that 79.3% percent of the images generated by CycleGAN are fake, while 20.7% are real photographs. We believe the blind audition could reflect the quality and fidelity of the generated images, so as to determine the performance of different models.

5 Conclusions

In this paper, we investigate the synthesis of traditional Chinese paintings of different styles. We propose CA-GAN to solve style transfer between Chinese line drawing and meticulous drawing & ink drawing. The main innovative ideas include decoupled content & attribute spaces and attention mechanism so as to synthesize vivid and higher-resolution traditional Chinese paintings. Besides, the utilization of MS-SSIM and Charbonnier losses ensures the generalization ability of the model and also enriches the diversity of the generated images.

As far as we know, there are few research results regarding to the generation of traditional Chinese paintings. We hope

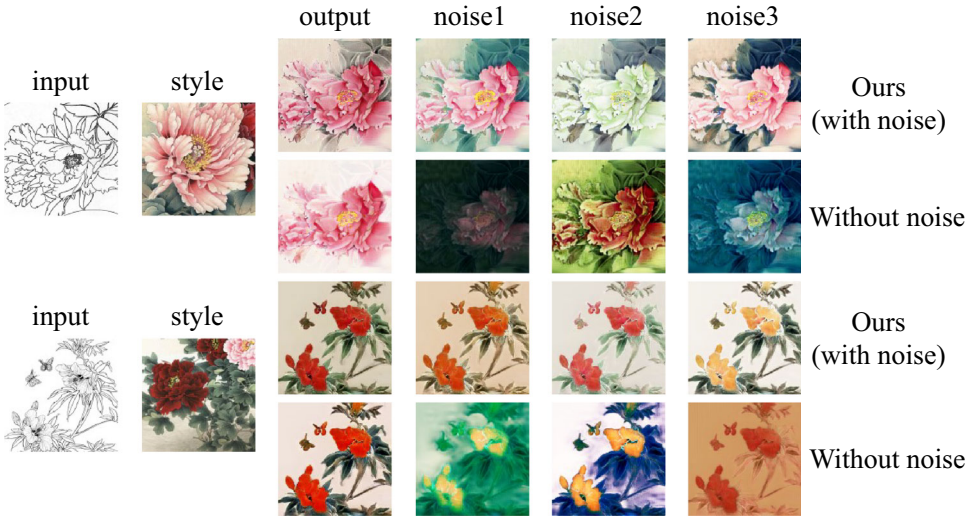

Fig. 5 Illustration of the function of Gaussian noise

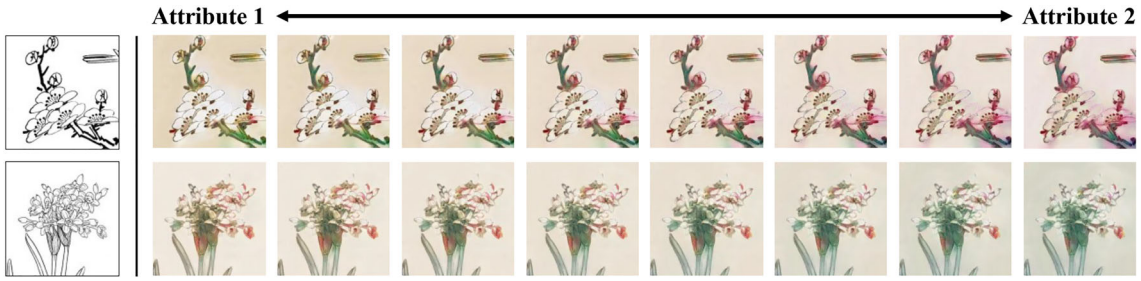

Fig. 6 Interpolation between two style attributes

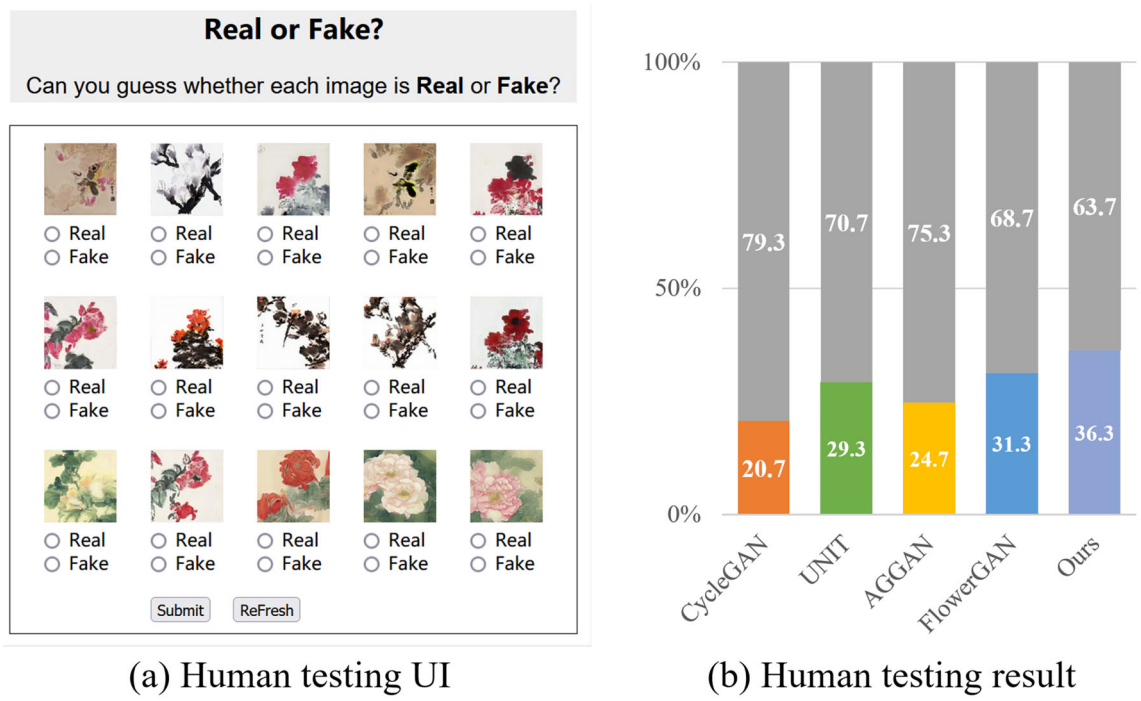

Fig. 7 Human testing UI and results on comparison among several methods

our work provide new ideas of image synthesis. In the future, we plan to carry out more relevant research in-depth on few-shot learning and generation of unique artistic styles.

Data availability The raw/processed data required to reproduce these findings will be shared once this paper has been accepted.

Declarations

Conflict of 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.

References

1. Jiang, Y., Wang, H.: Style transfer method based on deep learning for Chinese ink painting. In: Dang, N.H.T., Zhang, Y.-D., Tavares, J.M.R.S., Chen, B.-H. (eds.) Artificial Intelligence in Data and Big Data Processing, pp. 473–484. Springer, Cham (2022)
2. Zhang, J., Duan, Y., Gu, X.: Research on emotion analysis of Chinese literati painting images based on deep learning. Front. Psychol. (2021). https://doi.org/10.3389/fpsyg.2021.723325
3. Li, Z., Lin, S., Peng, Y.: Chinese painting style transfer system based on machine learning. In: 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), pp. 38–41 (2021). https://doi.org/10.1109/ICDSCA53499.2021.9650335
4. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp. 2414–2423. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.265
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014)
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8–13 2014, Montreal, Quebec, Canada, pp. 2672–2680 (2014). https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html
7. Lin, D., Wang, Y., Xu, G., Li, J.Y., Fu, K.: Transform a simple sketch to a Chinese painting by a multiscale deep neural network. Algorithms **11**, 4 (2018)
8. He, B., Gao, F., Ma, D., Shi, B., Duan, L.-Y.: Chipgan: a generative adversarial network for Chinese ink wash painting style transfer. In: Proceedings of the 26th ACM international conference on Multimedia (2018)
9. Yu, J., Luo, G., Peng, Q.: Image-based synthesis of Chinese landscape painting. J. Comput. Sci. Technol. **18**, 22–28 (2008)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015—18th International Conference Munich, Germany, October 5–9, 2015, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 1558–1566. JMLR.org (2016). http://proceedings.mlr.press/v48/larsen16.html
12. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: CVAE-GAN: fine-grained image generation through asymmetric training. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp. 2764–2773. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.299
13. Tang, H., Liu, H., Xu, D., Torr, P.H.S., Sebe, N.: Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. IEEE Trans. Neural Netw. Learn. Syst. (2021). https://doi.org/10.1109/TNNLS.2021.3105725
14. Tang, H., Xu, D., Sebe, N., Yan, Y.: Attention-guided generative adversarial networks for unsupervised image-to-image translation. In: International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14–19, 2019, pp. 1–8. IEEE (2019). https://doi.org/10.1109/IJCNN.2019.8851881
15. Kim, J., Kim, M., Kang, H., Lee, K.: U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net (2020). https://openreview.net/forum?id=BJlZ5ySKPH
16. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 3697–3707 (2018). https://proceedings.neurips.cc/paper/2018/hash/4e87337f366f72daa424dae11df0538c-Abstract.html
17. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-attention generative adversarial networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 7354–7363. PMLR (2019). http://proceedings.mlr.press/v97/zhang19d.html
18. Jiang, Y., Chang, S., Wang, Z.: Transgan: two transformers can make one strong GAN. 2021. CoRR arXiv:2102.07074
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, pp. 5998–6008. Long Beach, CA, USA (2017). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net (2021). https://openreview.net/forum?id=YicbFdNTTy
21. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net (2017). https://openreview.net/forum?id=Hk4_qw5xe

22. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, pp. 5767–5777. Long Beach, CA, USA (2017). https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html
23. Nowozin, S., Cseke, B., Tomioka, R.: f-gan: Training generative neural samplers using variational divergence minimization. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, pp. 271–279. Barcelona, Spain (2016). https://proceedings.neurips.cc/paper/2016/hash/cedebb6e872f539bef8c3f919874e9d7-Abstract.html
24. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp. 2813–2821. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.304
25. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, pp. 5967–5976. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.632
26. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp. 2242–2251. IEEE Computer Society (2017). https://doi.org/10.1109/ICCV.2017.244
27. Liu, M., Breuel, T.M., Kautz, J.: Unsupervised image-to-image translation networks. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, pp 700–708. Long Beach, CA, USA, (2017). https://proceedings.neurips.cc/paper/2017/hash/dc6a6489640ca02b0d42dabeb8e46bb7-Abstract.html
28. Huang, X., Liu, M., Belongie, S.J., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III. Lecture Notes in Computer Science, vol. 11207, pp. 179–196. Springer (2018). https://doi.org/10.1007/978-3-030-01219-9_11
29. Lee, H., Tseng, H., Huang, J., Singh, M., Yang, M.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision—ECCV 2018—15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11205, pp. 36–52. Springer (2018). https://doi.org/10.1007/978-3-030-01246-5_3
30. Lee, H., Tseng, H., Mao, Q., Huang, J., Lu, Y., Singh, M., Yang, M.: DRIT++: diverse image-to-image translation via disentangled representations. Int. J. Comput. Vis. **128**(10), 2402–2417 (2020). https://doi.org/10.1007/s11263-019-01284-z
31. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003, vol. 2, pp. 1398–14022 (2003). https://doi.org/10.1109/ACSSC.2003.1292216
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
33. Lai, W., Huang, J., Ahuja, N., Yang, M.: Fast and accurate image super-resolution with deep Laplacian pyramid networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(11), 2599–2613 (2019). https://doi.org/10.1109/TPAMI.2018.2865304
34. Fu, F., Lv, J., Tang, C., Li, M.: Multi-style Chinese art painting generation of flowers. IET Image Process. **15**(3), 746–762 (2021). https://doi.org/10.1049/ipr2.12059
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings (2015). arXiv:http://arxiv.org/abs/1412.6980
36. Barratt, S.T., Sharma, R.: A note on the inception score. 2018. CoRR arXiv:1801.01973
37. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, pp. 6626–6637. Long Beach, CA, USA (2017). https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

Zihan Chen received a B.Eng. degree from Sichuan University, China in 2021. He is now pursuing his master degree focusing on computer vision.

Yi Zhang received a B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China in 2004, and the Ph.D. degree from the National University of Singapore, Singapore, in 2010, respectively. He is currently an Associate Professor with the Department of Computer Science, Sichuan University, China. His current research interests include image processing, computer vision, machine learning, and robotics.