International Journal of Software and Informatics, ISSN 1673-7288
http://www.ijsi.org, ijsi@iscas.ac.cn, +86-10-62661048
IJSI, 2023, 13(2): 197–219, doi: 10.21655/ijsi.1673-7288.00314

International Journal of Software and Informatics

Research Article

Text-to-Chinese-painting Method Based on Multi-domain VQGAN

Zelong Sun (孙泽龙)[1], Guoxing Yang (杨国兴)[1], Jingyuan Wen (温静远)[1], Nanyi Fei (费楠益)[2], Zhiwu Lu (卢志武)[1], Jirong Wen (文继荣)[1]

[1] (Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China)
[2] (School of Information, Renmin University of China, Beijing 100872, China)
Corresponding author: Zhiwu Lu, luzhiwu@ruc.edu.cn

Abstract With the development of generative adversarial networks, synthesizing images from text descriptions has become an active research area. However, text descriptions used for image generation are often in English, and the generated objects are mostly faces, flowers, birds, etc. Few studies have been conducted on the generation of Chinese paintings with Chinese descriptions. The text-to-image task often requires a large number of labeled image–text pairs, which is expensive and boring. The advance of vision-language pre-training enables an image generation process guided by an optimized way, which significantly reduces the demand for annotated datasets and computational resources. In this paper, a multi-domain VQGAN model is proposed to generate Chinese paintings in multiple domains. Further, a vision-language pre-training model WenLan is used to calculate the distance loss between the generated images and the text descriptions. The semantic consistency between images and text is achieved by optimizing the hidden space variables as the input of multi-domain VQGAN. An ablation study is conducted to compare different variants of our multi-domain VQGAN in terms of the FID and R-precision metrics. We also conduct a user study to further show the effectiveness of our proposed model. The extensive results demonstrate that our proposed multi-domain VQGAN model outperforms all the competitors in terms of image quality and text-image semantic consistency.

Keywords Chinese painting generation; multi-domain generation; text-to-image generation

Citation Sun ZL, Yang GX, Wen JY, Fei NY, Lu ZW, Wen JR. Text-to-Chinese-painting method based on multi-domain VQGAN, *International Journal of Software and Informatics*, 2023, 13(2): 197–219. http://www.ijsi.org/1673-7288/314.htm

When we hear or read a story, corresponding pictures come to our minds involuntarily. It is so natural for human beings to combine the visual world with the linguistic world that we are often ignorant of this complex process. The mental faculty of conceiving human vision plays a pivotal role in the cognition process, such as memory, spatial imagination, and inference[1]. Inspired by the ways to visualize scenes described as text, the system that can grasp the relationships

This is the English version of the Chinese article “基于多域 VQGAN 的文本生成国画方法研究. 软件学报, 2023, 34(5): 2116–2133. DOI: 10.13328/j.cnki.jos.006769”
Funding items: National Natural Science Foundation of China (61976220, 61832017); Program for Outstanding Young Scientists at Universities or Colleges in Beijing (BJJWZYJH012019100020098)
Received 2022-04-16; Revised 2022-05-29; Accepted 2022-08-24; IJSI published online 2023-06-29

between vision and language and generate quality pictures corresponding to text descriptions is a milestone during the development of artificial intelligence.

There are many text-to-image models based on Deep Neural Networks (DNNs), and their major differences are in the ways to represent the semantic space of text and images, as well as the models or methods used to connect the semantic space of the two. Methods in the early stage, such as StackGAN[2], DMGAN[3], and AttnGAN[4], attempted to train a convolutional generator and directly predicted the pixels of images from the eigenvectors of a given text. However, these models had poor generalization ability and demonstrated poor performance in image quality and text–image matching when applied to generate common images.

Recently, DALL-E[5] and CogView[6] have reported good performance in the text-to-image field. To enable the representation of images like natural language, both adopt Transformer[7] and use vector quantification models similar to the Vector Quantized-Variational AutoEncoder (VQVAE)[8] and Vector Quantization Generative Adversarial Network (VQGAN)[9] for discrete representation of eigenvectors in the latent space of images. After that, they splice the representations of latent-space features of text and images and input them into Transformer to train the cross-modal text and image data in a unified framework. Despite the good generation effect, such large-scale models require hundreds of millions of pairs of text and image data for training, and the construction cost of such datasets is high.

The current text-to-image models mostly target a specific domain (human faces, flowers, birds, etc.) and can hardly generate images from multiple domains with just one model. In addition, the text used in image generation is mostly in English, and few studies focus on Chinese and Chinese paintings. Considering the above problems, a multi-domain Chinese painting generation method based on the vision-language pre-training model WenLan of Chinese and the generative model VQGAN is proposed. This method can use little data without text annotations to generate Chinese landscape paintings from Chinese poetry according to different domains. The overall framework of the proposed model is shown in Figure 1, which involves two stages of generation. In the first stage, the network structure of multi-domain VQGAN is designed, which can receive the image input from multiple domains simultaneously and generate images from different domains. In the second stage, the encoder of the trained multi-domain VQGAN is removed, and random noise is utilized to generate images as the input of WenLan.

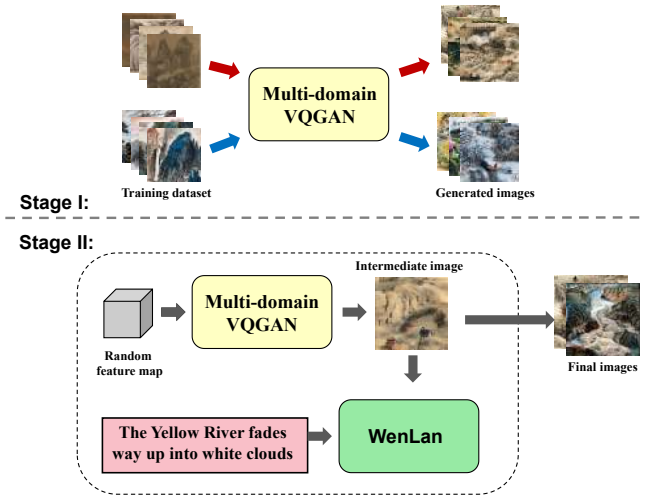


Figure 1 Framework of text-to-Chinese painting model based on multi-domain VQGAN

Then, WenLan encodes text and images and extracts their features, and these features undergo similarity calculation to lead to the loss function. The random noise is constantly updated to make the generated images closer to the text descriptions.

The major contributions of this work are as follows. (1) A multi-domain text-to-image model is proposed, which is capable of generating images from multiple domains. In addition, the training progress does not require datasets with text annotations, which avoids the problem that large-scale labeled text and image pairs are difficult to acquire. (2) The proposed model takes Chinese descriptions as the text input and can generate high-resolution Chinese paintings using ancient Chinese poetry.

Section 1 introduces the methods and research status of text-to-image models, and Section 2 specifies the required basic knowledge, including code dictionaries, Generative Adversarial Networks (GANs), and multi-modal pretraining models. Sections 3 and 4 present the built model for generating multi-domain Chinese paintings based on multi-domain VQGAN and WenLan. Section 5 compares the images generated by the proposed model and the latest model by the ablation experiment and user investigation experiment to verify the effectiveness of the proposed model. Finally, a summary is given.

1 Related Work

1.1 Text-to-image generation

The text-to-image methods could date back to the early stage of deep generative models. At that time, Mansimov *et al*. added text into DRAW[12]. After that, due to the outstanding performance of GANs in image synthesis technology[13], methods based on GANs began to dominate text-to-image tasks.

In 2016, Reed *et al*.[14] attempted to apply GANs to text-to-image tasks for the first time. In their model, text is input into the generator as conditions to constrain the generation of images. The discriminator needs to discriminate two situations, namely real data and a mismatch between generated images and text descriptions. Eventually, images with a resolution of 64×64 can be generated. To upgrade the quality of the generated images, Reed *et al*. proposed the GAWWN[15] model in the same year. On the basis of their previous model, this model uses specified key points to restrain the parts with different objects in an image and thus make the text correspond to image details. In the end, images with a resolution of 128×128 can be obtained. Hu *et al*.[16] put forward a text-to-image network based on single-stage GANs. In this model, the channel-pixel attention module is added to the generator, and the global text representation and local word embedding techniques are employed to provide fine-grained information for discrimination, which enables a generator and a discriminator to generate quality images. In contrast, StackGAN[2] regards the generation process as the refinement process of schematic diagrams and expands a single generator and discriminator pair in traditional methods to two pairs of generators and discriminators. The generation process of StackGAN is composed of two stages. In the first stage, rough images with a resolution of 64×64 are generated under given random noise vectors and text vectors. In the second stage, the original image and text vectors are input into the second generator to finally output images with a resolution of 256×256. In the two stages, each discriminator is trained to discriminate between matched and mismatched image-text pairs. StackGAN++[17] further improves the system structure through an end-to-end framework. On the basis of StackGAN, the stack structure is improved to be a tree structure, and three pairs of generators and discriminators are jointly trained. In addition, the model uses color consistency regularization to minimize the differences in the average and covariance of the pixels of different scales. FusedGAN[18] draws upon the idea of simultaneously training conditional and unconditional distributions, which is comprised of two generators, one for

conditional image generation and the other for unconditional image generation, and they share a common latent space. To address the multi-generator problem in one model, HDGAN[19] proposes an accompanying hierarchically-nested adversarial objective, which regularizes the generation of low-resolution images at the mid-level to enable the generator to capture complex image information. During the generation process, each mid-level is nested with a discriminator to decide whether a generated image is true or false and describe the semantic correlations of text. Similarly, PPAN[20] only uses one generator with a pyramid framework[21] and three different discriminators. The generator combines the features with a low resolution and strong semantics and features with a high resolution and weak semantics through a lateral connection path from the bottom up. AttnGAN[4], on the basis of StackGAN++, fuses the attention mechanism into the multi-stage generation process. As the attention mechanism allows the network to synthesize fine-grained details according to relevant word and global sentence vectors, the network can better align images and text. MirrorGAN[22], following the idea of "learning text-to-image generation by redescription", proposes a text-to-image-to-text architecture. After image generation, the matching text is reversely generated, and the loss between reversely generated text and original text is calculated, which enables better learning of the semantic consistency between generated image and text. DMGAN[3] replaces the attention mechanism in AttnGAN with a dynamic memory model. In this way, the model can generate more vivid images. CAE-GAN[23] translates and aligns text information and visual information by a cross-attention encoder to capture the cross-modal mapping relationship between text and image information for higher fidelity of generated images and higher matching degrees with input text descriptions.

In addition to the text-to-image models based on GANs, the latest representatives DALL-E[5] and CogView[6] use the Transformer generation structure. This structure has at least billions of parameters and requires massive high-quality text-image pairs to pre-train the Transformer. During pre-training, the models serialize the images by using tokenizer with VQVAE[8] and splice the serialized text as the input of the Transformer model to generate images. In the final stage of processing text-to-image tasks, the models sort the generated images with a calculated Caption Score and select the image with the highest matching degree with the text as the final result. Despite the good effect, such large-scale models should be trained with hundreds of millions of text–image pairs, and the construction cost of such datasets is drastically high. Moreover, the computing resources employed for model training are considerable, which can hardly be satisfied by general research institutes.

1.2 Optimization-based GAN inversion

For a trained unconditional GAN model, GAN inversion attempts to encode the original image x into the latent-space representation z^*, and in this way, the image generated by z^* can be visually similar to x. GAN inversion is realized mainly in three methods, i.e., learning-based method, optimization-based method, and a mixture of the two. The optimization-based method is similar to the method used in this paper, and thus, this section mainly introduces the relevant contents of this method. Figure 2 shows the general process of this method.

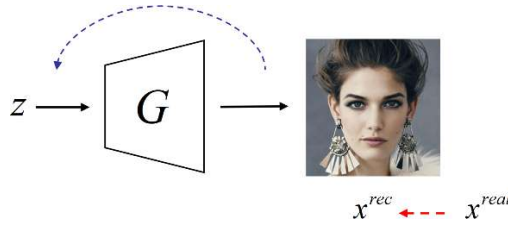


Figure 2 General process of optimization-based GAN inversion

The existing optimization-based GAN inversion methods generally optimize implicit vectors to construct the target image, namely,

$$z^* = \underset{z}{\operatorname{argmin}}\, L(G(z), x) \tag{1}$$

where L is a distance measure function, and G is the pre-trained generator. Three issues should be tackled by the optimization-based GAN inversion methods, namely, determining ways to optimize the vector z, ways to solve the local minimum problem, and ways to initialize the vector z.

In previous work, Creswell *et al*.[24] chose to optimize the implicit vector z by the gradient descent method, and in the optimization process, inversion was performed on a batch of image samples at once. This method can not only offset the impact of batch normalization in the inversion process but also allow parallel inversion of multiple image samples. The gradient descent method was also used in Image2StyleGAN[25]. A given image is mapped to the extended latent space W+ of the pre-trained StyleGAN[26], and the structure of the latent space is further explained through three basic operations, i.e., linear interpolation, crossover, and addition of vectors and proportional difference vectors. On this basis, Image2StyleGAN++[27] innovatively adds noise to improve the embedding quality and generate higher-quality embedded images. Moreover, it optimizes the embedding first and then the noise to obtain a higher peak signal-to-noise ratio than that of the simultaneous optimization of the two. Voynov *et al*.[28] used Jacobian decomposition to analyze the latent space of the pre-trained GAN model and introduced a lightweight method that could identify several directions at the same time to reduce the high cost brought about by the calculation of a large number of Jacobian matrices.

The local minimum problem is generally solved by two types of optimizers: one is gradient-based, such as ADAM[29] and L-BFGS[30], and the other is gradient-free, such as covariance matrix adaptation (CMA)[31]. For example, Image2StyleGAN[25] uses the ADAM optimizer, while Zhu *et al*.[32] used the L-BFGS scheme in their work. Huh *et al*.[33] used the default optimization hyperparameters to test various gradient-free optimization methods, and they found that CMA and its variant BasinCMA register the best performance during the inversion of images from challenging data sets to the latent space of StyleGAN.

Since Eq. (1) is usually non-convex, the reconstruction quality usually depends strongly on the initialization of z. Image2StyleGAN[25] analyzes two initialization methods, namely, random initialization and the average implicit vector method, and finds that the results obtained by random initialization are better. However, a large amount of random initialization may be required for stable reconstruction[34], which makes real-time processing impossible. Some studies[34, 35] have proposed using encoders to obtain better initialization vectors.

The optimization-based GAN inversion methods and this paper both optimize the vector or feature map of the latent space to influence the finally generated image. As the iterative algorithm is used for optimization, there is a certain similarity between the two. It should be noted that the traditional GAN inversion methods do not introduce text information while this paper focuses on optimizing the feature map of the latent space through text to obtain the generated image similar to the text semantics.

2 Basics

2.1 Codebook

Learning the effective representation of images without supervision has always been a key challenge of machine learning. An autoencoder is a powerful generative model. Its principle is to encode the data x through the encoder and map it to the vector z in latent space, that is,

$z = encoder(x)$. After that, the data x is reconstructed into an image from the implicit vector z in the latent space through a decoder, that is, $x' = decoder(z)$. When the reconstruction loss between original image and reconstructed image is small, it can be considered that the network has learned the effective latent-space representation of the image.

On the basis of the autoencoder, a variational autoencoder[36] adds a constraint to ensure z satisfies the isotropic Gaussian distribution, which is called a priori distribution. Hence, after the training of the variational autoencoder, random sampling can be carried out in this a priori distribution to obtain a random latent-space noise, which is then decoded by a decoder to produce a random picture. Each dimension of the implicit variable z of the variational autoencoder is a continuous value.

Oord *et al.*[8] proposed the VQVAE method to learn the discrete representation of an image and used a convolutional structure for autoregression modeling of its distribution. VQGAN[9] introduces adversarial training loss to strengthen the learning of code dictionaries rich in perceptions. Each dimension of the implicit vector z of VQGAN is a discrete integer, which can make effective use of the latent space. Learning a codebook is the key to the discretization of z. Assuming that the codebook is composed of K D-dimensional vectors, an $H \times W \times D$ feature map z' can be obtained after the image is processed by the decoder. Then, the $H \times W$ D-dimensional vectors look up the codebook for the vectors with the closest distances for substitution. In this way, the quantified z_q is obtained, which is input into the decoder to generate an image.

2.2 GAN

Goodfellow *et al.*[37] first proposed GANs in 2014. The original intention of learning GANs is to generate data that does not exist in the real world. Compared with the traditional neural network models, a GAN is a new unsupervised architecture. It includes two independent networks, namely, generator and discriminator, which serve as the adversarial target of each other. As shown in Figure 3, the generator $G(z)$ receives the noise randomly sampled from a priori noise distribution as the input and generates an image x_g. The input of the discriminator is a real picture x_r or x_g, and its output is the probability that the input picture is a real picture.

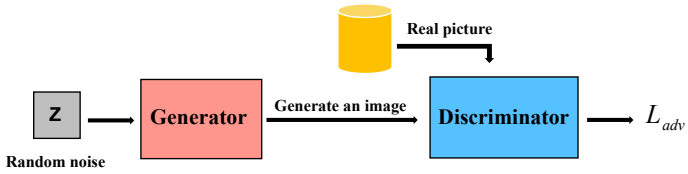


Figure 3 Structure of GAN

The training process of a GAN can be regarded as the adversarial process of the two networks. The discriminator is trained to tell whether the input is a real picture or a generated image, while the generator is trained to capture the real data distribution and generate an image as real as possible to fool the discriminator. For example, the training process of Goodfellow's GAN can be defined as the minimax game process of two networks with regard to the loss function $V(D, G)$. The training goal of the discriminator is to maximize the allocated possibility of the correct category while the training goal of the generator is to minimize the probability that the generated images are judged as false by the discriminator. The loss function is as follows:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x|y)] + \mathbb{E}_{x \sim P_z}[\log(1 - D(G(z|y)))] \quad (2)$$

2.3 Vision-language pre-training

In recent years, artificial intelligence has made great progress in both computer vision and natural language processing. Multimodal deep learning that integrates the two has also attracted more and more attention. The goal of vision-language pre-training is to align data from different modes and transfer the learned knowledge to various downstream tasks. The vision-language pre-training model can be divided into two types according to its framework: single tower and double tower.

The single-tower network treats the input of different modes such as images and text as the same and fuses them in the same model. Its representatives include UNITER[38], Oscar[39], M6[40], VisualBERT[41], Unicoder-VL[42], VL-BERT[43], and other models. They use a feature fusion module (such as Transformer[7]) to obtain the embedding of image–text pairs. Some single-tower models also use object detectors to detect image regions and match these regions with the corresponding words. As a representative of the single-tower model, UNITER conducts joint training of Masked Language Modeling (MLM), Masked Region Modeling (MRM), and Image-Text Matching (ITM) for 5.6 million image-text pairs to learn the common image-text representations. Oscar[39] uses Fast R-CNN[44] to associate the detected object label with the words in the text. However, the existing single-tower structure usually assumes that there is a strong semantic correlation between text and image modes, and the cross-modal interaction between image–text pairs should be simulated in a certain manner. However, this assumption is often invalid in real scenarios[10]. In addition, the single-tower model requires huge computational costs at the reasoning stage.

In contrast, the multimodal pre-training model with a double-tower structure processes the input of different modes separately and then carries out cross-fusion. It uses separate text and image encoders to encode the text and image before image-text matching. This mode has higher retrieval efficiency, but due to the lack of deeper image-text interactions, it usually can only achieve suboptimal performance. In recent work, LigningDot[45] meets this challenge by redesigning the target detection process, while CLIP[46], ALIGN[47], WenLan 1.0[10], and WenLan 2.0[11] quit the object detector with a high computational cost and use the cross-modal comparative learning task for model training.

3 Chinese Painting Generation Based on Multi-domain VQGAN

The vision-language pre-training model has a strong semantic understanding ability of text and images and can bridge the text and images while the GAN model can generate realistic images according to random low-dimensional noise. Therefore, this paper combines multimodal pre-training with the GAN model and uses the vision-language pre-training model to guide the image generation process of the GAN model. Specifically, the method involves two stages: image reconstruction by multi-domain VQGAN and text-guided image generation through WenLan. This paper introduces the two processes in two sections, and this section mainly presents the design and implementation of multi-domain VQGAN in the first stage.

In the previous text-to-image work, many models are faced with the following problems: (1) massive image-text pair data should be collected as datasets, which is costly; (2) the model can only take effect in a specific limited domain and is ineffective in multi-domain image generation, and it can hardly generate high resolution images. Therefore, we use the improved multi-domain VQGAN to generate images. The merit of such an effort is that we do not need matched pairs of images and text in the training process but only need separate images. Hence, the dataset construction overhead is greatly relieved. The improved multi-domain VQGAN can receive training datasets from multiple domains and generate images of multiple domains, and the image quality and resolution will be improved.

The purpose of GANs is to enable the data distribution of generated samples to fit that of real samples through adversarial training of the generator and discriminator and thereby obtain data that can be confused with real data[48]. Similar to VQVAE[8], VQGAN can learn an effective codebook. During image generation, the feature map is first initialized randomly and quantified in the codebook. After that, the decoder is applied. We can generate a more realistic image. However, even for the same kind of art form, there are many different classes and styles. For example, traditional Chinese paintings can be divided into ink wash painting, colored painting, paintings drawn with plain lines, etc., which are defined as "domain" in this paper. If the images belonging to the same art form but different domains are mixed to train the generative model, we cannot control the domain of the generated image. If the images of different domains are taken as the datasets of different models, there will be too many final models and model overfitting caused by the small datasets. Therefore, inspired by StarGAN v2[49], we adopt a branching method similar to it to share and branch the three important parts of VQGAN, i.e., encoder, codebook, and decoder, to different degrees so that the same model can be used to generate Chinese paintings from different domains.

Specifically, the model structure of VQGAN is divided into encoder, codebook, decoder, and discriminator. The multi-domain VQGAN proposed in this paper is composed of a k-domain encoder, a k-domain codebook, a k-domain decoder, and a discriminator. For $k = 1$, the multi-domain VQGAN is the same as the original VQGAN model. For $k > 1$, the down-sampling network will be shared by multiple domains for the k-domain encoder, and the rest will be the unique output branch of each domain. Like the encoder, the decoder will share the up-sampling network, and the rest will be the unique input branch of each domain. The codebook as a whole is independently owned by each domain. It is found that appropriate sharing of a certain number of networks in multi-domain training can mutually reinforce the quality, diversity, and relevance of text. In Section 5, we compare the generation effects of different combinations of the three components in detail.

The network structure and training process of multi-domain VQGAN are shown in Figure 4. For example, when $k = 2$, the domains of images are distinguished by simple labeling during the construction of the image datasets, and then the data of the two domains are mixed. During training, the same batch of data may contain data from two domains. After they pass the shared up-sampling network, they select the branch to enter according to the label of the domain to obtain the final high-resolution reconstructed images through the unified up-sampling process.

4 WenLan-guided Text-to-Chinese Painting

Although the multi-domain VQGAN model can generate images of different styles after training, its received input is random low-dimensional noise, and the generated results are also completely random images. Therefore, this paper introduces the Chinese multimodal pre-training model WenLan to receive Chinese text input to guide image generation process with text information.

As the largest Chinese multimodal general pre-training model at present, WenLan has achieved excellent performance in both accuracy and retrieval speed in the mutual retrieval task of images and text. It has a strong multimodal semantic understanding ability and can well bridge Chinese text and images. Its proposed BriVL model uses a double-tower structure as model architecture, and comparative learning is introduced into the double-tower structure of BriVL. WenLan uses independent language and visual encoders to extract the feature vectors of language and visual input, respectively, and trains and aligns the two vectors in the comparative learning module. The double-tower structure allows us to replace the encoder module with the latest single-mode pre-training model, thus enhancing the effectiveness of the model. For a

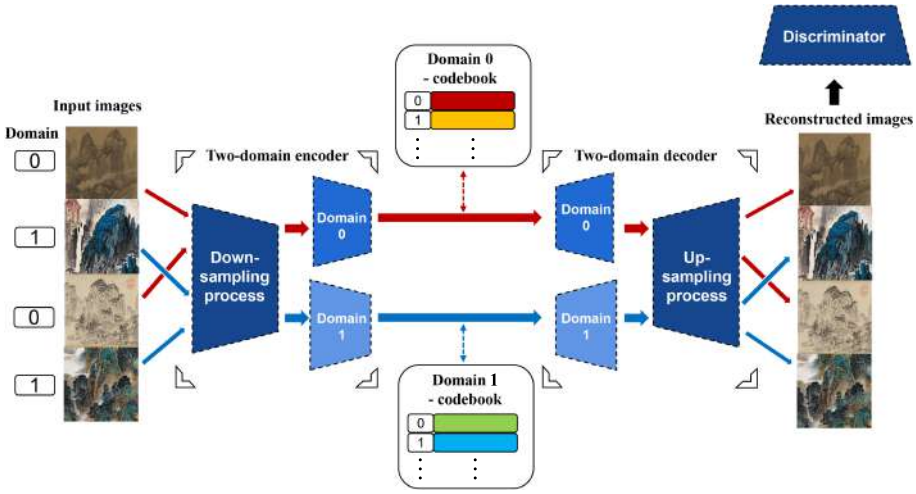


Figure 4 Network structure of multi-domain VQGAN

given text–image pair, BriVL uses both the visual mode and the language mode to construct the negative samples of the pair and expands the number of negative samples according to the idea of MoCo, thus improving the effectiveness of the neural network. Due to WenLan's double-tower structure, the pre-trained WenLan model can extract features from images and text separately, which is convenient for actual deployment and use. In addition, WenLan relaxes the strong data association assumption between multiple modes, and thus the abstract and flexible semantic associations can be used. This assumption of a "weak correlation" is more common in real life, which makes the generalization ability of the model stronger and also makes it possible for this paper to use more freehand and abstract poetry to generate images.

Up to now, WenLan has released two versions. Compared with WenLan 1.0[10], which uses 30 million image–text pairs to train its dataset, WenLan 2.0[11] uses 650 million weakly correlated image–text pairs from the Internet as its dataset, which has a stronger generalization ability. Additionally, WenLan 2.0 removes the target detector used in WenLan 1.0 which is an image–text matching model independent of the object detection results, thus reducing the computational overhead. Hence, it is more suitable for applications in the actual industry. Practical applications show that WenLan 2.0 has an intuitive understanding ability for some abstract concepts, such as "nature", which is understood as vast vegetation; for "time", its figurative understanding is a clock. Its understanding of some proverbs and phrases is also appropriate. Therefore, this paper uses WenLan 2.0 to guide the generation process of VQGAN, as shown in Figure 5. WenLan 2.0 is composed of an image encoder and a text encoder, where the image encoder takes EfficientNet[50] as the visual backbone network, and the text encoder takes RoBERTa-Large[51] as the textual backbone network. Upon the output of the above-mentioned backbone model, BriVL stacks four layers of Transformer[7] to obtain 2560-dimensional visual and text features and applies the loss function InfoNCE[52] to align the text features with the image features.

In image generation, a feature map is first initialized at random, and then the codebook of the corresponding domain is used for quantification according to the input domain label to obtain the quantified feature map. After that, a random image is generated when it is processed by the decoder of the corresponding domain, and the generated image is input into the image encoder of the WenLan model to obtain the image features. At the same time, the input text is input into the text encoder of the WenLan model to obtain the text features. Finally, the

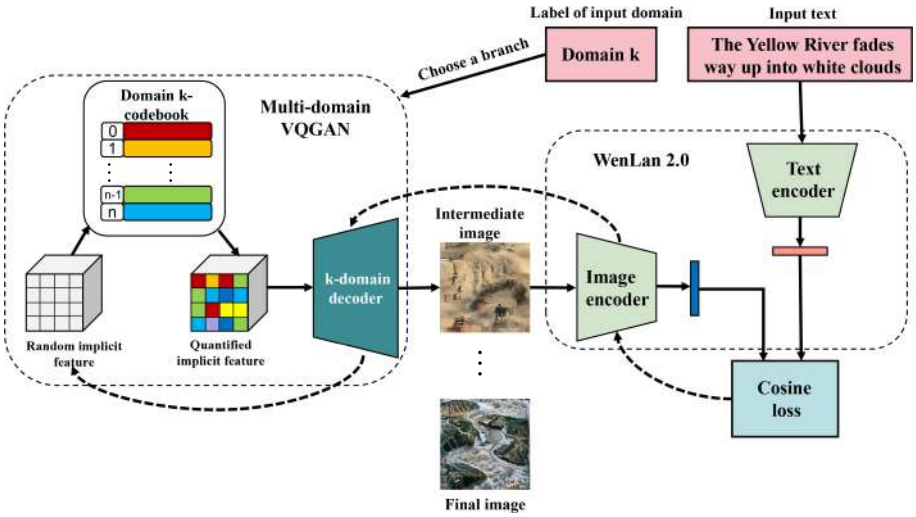


Figure 5 WenLan-guided multi-domain VQGAN for image generation

similarity between image features and text features is calculated as the loss function, and the random low-dimensional noise input into the VQGAN is continuously updated through gradient backpropagation to minimize the loss. In this way, the semantics of the generated image is gradually made consistent with the semantics of the text. A domain is taken as an example for explanation, as shown in Figure 5. When WenLan 2.0 guides the image generation process of VQGAN, the forward propagation process is as follows. (1) The feature map of the latent space is randomly initialized and quantified with the codebook of the corresponding domain to obtain the quantified feature map. (2) The quantified feature of the latent space passes through the decoder of the corresponding domain to generate an image. (3) The feature vector of the generated image is obtained through the image encoder of WenLan 2.0, and the feature vector of the input text is obtained through the text encoder of WenLan 2.0. (4) The cosine similarity between image feature vector and text feature vector is calculated, and the distance between them is measured as a loss function. In the back propagation process (dotted arrows), the gradient propagates through WenLan 2.0 and VQGAN to the random feature map of the latent space and updates it through the gradient descent method. In this way, the semantics of the generated image is closer to the semantics of the input text description.

5 Experimental Analysis

5.1 Experimental data

The image datasets used in this paper mainly include ink wash paintings and colored paintings.

Colored paintings. The dataset of colored paintings is self-constructed. A total of 7,000 original colored paintings through multiple channels, such as the Internet and WeChat subscriptions, were collected and processed, which took a lot of time. First, non-landscape and low-quality paintings were manually filtered out. Second, the centers of paintings were manually cut out, and the prefaces and postscripts were removed. Third, the centers of the paintings were cut, and the textual information such as inscriptions was removed. Fourth, each painting was cut into multiple square paintings according to its shortest edge, and its resolution was adjusted to 512×512. In a word, the final number of colored paintings is 2,925. Figure 6 shows the colored painting dataset constructed in this paper.

Ink wash painting. The dataset of ink wash paintings contains 2,811 images with a resolution of 512×512. The source is mainly divided into two parts. One part is from the dataset published by Xue[53], whose data comes from open-access museums and galleries including the Freer Gallery of Art of the Smithsonian's National Museum of Asian Art, the Metropolitan Museum of Art, the Princeton University Art Museum, and the Harvard Art Museums. The dataset has been manually filtered and cut by the creator and was filtered again in this study, and finally, a total of 1,979 paintings with a resolution of 512×512 were obtained. Another part of the ink wash painting dataset is the high-resolution paintings we collected on the Internet. The paintings were mainly created in the Tang, Ming, and Qing dynasties. After manual filtering, textual information removal, and cutting, a total of 832 images with a resolution of 512×512 were obtained. Figure 7 shows some samples of the ink wash painting dataset used in this paper. The first row of paintings is from the public data set, and the second row is from the dataset constructed in this paper.

Figure 6 Samples of colored painting dataset

Figure 7 Samples of ink wash painting dataset

5.2 Evaluation indexes

A quantitative analysis of image quality and image-text alignment was made. In addition, a survey of users' subjective evaluation was conducted to obtain users' evaluations for the generated results, which was the qualitative analysis of this work.

Analysis of image quality. The FID index[54] was employed to evaluate the image quality. FID is a score calculation method based on the Inception network, an image classifier trained by the ImageNet dataset. In the calculation of the FID index, the last pooling layer of the pre-training Inception v3[55] was removed to extract the image features, and then the distance between the real image distribution and the generated image distribution was measured. Assuming that the obtained image features follow the multi-dimensional Gaussian distribution, the distance between the two distributions can be calculated by the mean and covariance matrices. Eq. (3) gives the FID between the real data and the generated data, where $(\mu_r, \sum r)$ is the mean and covariance of the distribution of features extracted from the real image, and $(\mu_g, \sum g)$ is the mean and covariance of the distribution of features extracted from the generated image.

$$FID = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\sum r + \sum g - 2\left(\sum r \sum g\right)^{1/2}\right) \tag{3}$$

FID represents the distance between the feature vector of the generated image and the feature vector of the real image. A closer distance corresponds to a smaller FID and a better effect of the generative model. In other words, the generated image has higher definition and rich diversity.

Image and text alignment. The capability to generate a real image is only one aspect of the evaluation of the text-to-image model. Another important aspect is to evaluate the semantic alignment between the text description and the generated image. Therefore, the R-precision index[4] was adopted in this paper, whose principle is to sort the retrieval results between the features of the generated image and the text features and thus measure the visual semantic similarity between the text description and the generated image. The cosine similarity between the feature of each generated image and each text feature in the text set was calculated, and the text were sorted in descending order of similarity. If the real text used to generate the image was on the top R, the generated images would be considered successful.

Users study. The FID index can make a quantitative evaluation of image quality and image diversity, while the R-precision index can well evaluate the similarity between images and text. However, FID and R-precision are not comprehensive in the evaluation dimension. Therefore, the subjective evaluation by users was introduced in this study, which could present qualitative evaluations of experimental results in multiple dimensions such as object fidelity, image interpretability, and common sense[55]. The subjective evaluation of users could be obtained as follows: the generative model was used to generate a certain number of images from a certain number of randomly sampled text, and then the text and the corresponding generated images were provided to users to select the best image or sort them.

5.3 Ablation experiment

Multiple combination methods were designed to fully explore the impact of the k-domain encoder, k-domain codebook, and k-domain decoder on the final generation effect, and they were trained on two datasets with the same parameters. FID and R-precision indexes were obtained through tests.

Experimental design. Eight network structures were designed in this section: the original VQGAN structure, namely that the encoder, codebook, and decoder were all one-domain; only one of the encoder, codebook, and decoder was replaced with a two-domain structure, and the rest were one-domain; two of the encoder, codebook, and decoder were replaced with two-domain structures, and the left one was kept in the original VQGAN structure; the encoder, codebook, and decoder were all replaced with a two-domain structure. See Table 1 for the specific deployment of the model.

The ink wash painting and colored painting, two domains, were taken as the dataset to train the model, and the following indexes were all calculated in the two domains separately. Two hundred lines from Chinese poetry were collected, one hundred of which were randomly selected for image generation. Five images were generated for each line, and thus 500 images were obtained to form a test set of generated images. Samples of the line set are displayed in Table 2.

Table 1 Eight model combinations

Model	Encoder with a two-domain structure	Codebook with a two-domain structure	Decoder with a two-domain structure
Original VQGAN model	×	×	×
Two-domain encoder	√	×	×
Two-domain codebook	×	√	×
Two-domain decoder	×	×	√
Two-domain encoder and decoder	√	×	√
Two-domain encoder and codebook	√	√	×
Two-domain codebook and decoder	×	√	√
Complete model	√	√	√

Table 2 Lines of poems used for image generation

No.	Line of poems
Text 0	The sun beyond the mountain bows. The Yellow River seawards flows.
Text 1	But where's this Fragrance Temple, where's it, though? I've soon been plunged in the maze of cloudy peaks!
Text 2	At the eastern city gate of Luntai I shall see you off. The road ahead along Tianshan Mountains is heavy with snow.
Text 3	It's just like a whole river full of eastward flow in spring.

***FID* calculation**. To calculate the FID index, the real image set used the training set, and the generated image set used the generated 500 images. The features of the two data sets were extracted by the Inception-v3 network, and each image had a 2048-dimensional vector. After that, the FID was calculated by Eq. (3).

Retrieval recall experiment. The generated images were utilized to recall their corresponding text descriptions during the calculation of the R-precision. Specifically, the generated 500 images were used to retrieve the 200 lines of poems. First, WenLan 1.0 and WenLan 2.0 were used to extract the features of all the generated images and all the text separately. Then, the cosine similarity between the feature vectors of each generated image and all the text features was calculated. Finally, the candidate text descriptions of each image were ranked in descending order of similarity, and the most similar text was used to calculate the R-precision index. In other words, only when the real text corresponding to the image ranks first in the similarity ranking can the text recall be considered successful.

The models in this paper all operated for 300 rounds under the same setting of hyperparameters. The parameter quantity of each part of the model is shown in Table 3. It can be seen that when the complete model of two domains is applied, the parameter quantity is saved by about 24M compared with that of the two original VQGAN models. When the number of domains increases, the saving effect of the parameter quantity becomes more significant, which greatly reduces the demand for hardware and improves the training efficiency.

The encoder, codebook, and decoder of the two domains were trained separately according to the eight combinations shown in Table 1, and the FID and R-precision indexes were calculated. See Table 4 for the specific experimental results. When the datasets of the two domains are used to train the original VQGAN model separately, the FID value is low, which means that

Table 3 Parameter quantity for each part of the model

Domain	Encoder (M)	Codebook (k)	Decoder (M)
One-domain	29.3	262	42.4
Two-domain	41.1	524	54.3

the quality of the generated image is good. Under WenLan 1.0 or WenLan 2.0, the R-precision index is almost the lowest of all models; however, when a two-domain encoder or a two-domain decoder is added (Rows 2 and 4 in Table 4), the R-precision index is significantly improved. This shows that during the joint training of the datasets of the two domains, the datasets of the two domains can help each other and enrich their respective image objects thanks to the shared network when a two-domain encoder or a two-domain decoder is used. For example, when the input text is the Yellow River, if there is no river in the dataset of colored paintings, the image of the Yellow River cannot be generated by the original VQGAN model. Once the two datasets are jointly trained due to the partially shared network, the Yellow River or river information in ink wash paintings can be used for reference to improve the semantic consistency between the generated image and the input text in the two domains. However, when a two-domain encoder or two-domain decoder is used alone, the distribution of the two datasets affects each other because the proportion of the shared network is too large (for example, in the case of a two-domain encoder, the encoder's down-sampling network, codebook, and decoder will be shared), which results in a high FID index. When the two-domain codebook is used alone (Row 3 of Table 4), the FID value is the highest among the eight models, and the R-precision index is the lowest among the seven models that use the two-domain components. This indicates that the two datasets can hardly help each other when the two-domain codebook is used alone, which leads to the confusing distribution of each domain. When only the two-domain decoder is missing (Row 6 of Table 4), the FID value is large, which means that the decoder plays a pivotal role in distinguishing the distributions of different domains when generating images. In Rows 2–4 of Table 4, only the two-domain decoder is used, and the FID value is lower than that under only a two-domain encoder or a two-domain decoder, which also demonstrates the above conclusion. In the fifth row of Table 4, only the two-domain codebook is missing, which leads to a low R-precision index. This means that each domain has an independent codebook during image generation, which indeed helps improve the semantic consistency between text and images. When the complete model is applied, namely, the network structure shown in Figure 4, both the FID value and the R-precision index have achieved optimal results. This indicates that the datasets of the two domains can be well combined for training to improve the generation quality and effect of each domain.

Table 4 Comparison of experimental results under different model combinations

Model	Ink wash painting			Colored painting		
	FID	R-precision W2	R-precision W1	FID	R-precision W2	R-precision W1
Original VQGAN model	138.845	0.656	0.162	88.226	0.450	0.104
Two-domain encoder	160.277	0.812	0.202	93.120	0.788	0.202
Two-domain codebook	164.789	0.660	0.168	105.573	0.684	0.158
Two-domain decoder	154.890	0.780	0.168	89.793	0.768	0.194
Two-domain encoder and decoder	139.007	0.668	0.158	90.368	0.754	0.178
Two-domain encoder and codebook	160.540	0.758	0.230	96.323	0.796	0.220
Two-domain codebook and decoder	145.027	0.778	0.210	90.946	0.778	0.178
Complete model	**136.032**	**0.872**	**0.236**	**87.780**	**0.818**	**0.238**

Colored paintings were taken as examples for a more intuitive comparison of the impact of different combinations of the two-domain encoder, codebook, and decoder on the generated results. Figure 8 shows the results generated by the lines in Table 2 under different models. It

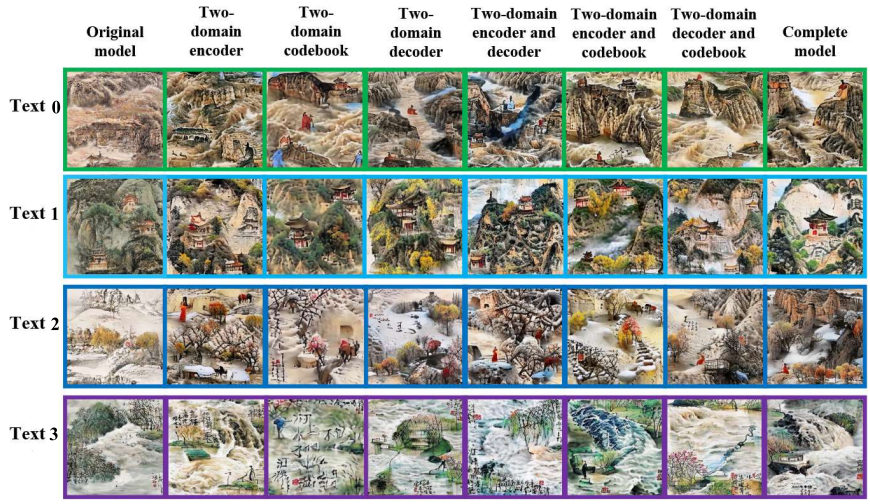


Figure 8 Generated images of multi-domain VQGAN under different branches

can be seen that the quality of the images generated by the original VQGAN model (the first column of Figure 8) is good, but the quality of those generated by the complete multi-domain VQGAN model is better in terms of color and object richness. In the case of Text 0, i.e., "The sun beyond the mountain bows. The Yellow River seawards flows", and Text 3, i.e., "It's just like a whole river full of eastward flow in spring", "flow in spring" and "Yellow River" in the images generated by the complete VQGAN model are clearer. The surge of the Yellow River and the momentum of the eastward flow in spring can be felt more deeply, and the descriptions of waves and water patterns are more obvious and detailed. In the case of text 0 and 2, the complete VQGAN model can also give more detailed descriptions of "(the sun) bowing beyond the mountain" and "Tianshan Mountains". However, in the images generated by the original model (Rows 1 and 3 of the first column in Figure 8), the shape of the mountains can hardly be seen. In the images generated by the models with some two-domain components (Columns 2–7 in Figure 8), the generation effect is the worst when only the two-domain codebook is used (Column 3 in Figure 8) while the generation effect is greatly improved under the two-domain encoder and the two-domain decoder (Column 5 in Figure 8). The result conforms to the above conclusions obtained from Table 4.

5.4 Users study

The subjective evaluation of users is introduced in this section to better evaluate the effect of the models using lines of poems to generate multi-domain traditional Chinese paintings. Sixty lines of poems were randomly selected from each of the two domains of ink wash paintings and colored paintings. For each line, eight models in Table 1 were employed to generate an image separately. The serial numbers of the generated images were mixed up to ensure the fairness of the experiment. Seven volunteers chose (1) the image closest to a real picture and (2) the image most relevant to the text description from the eight generated images. The experimental results are shown in Table 5. It can be seen that the generation effect of the complete model is better than that of other models; the effect of the original VQGAN model and that only using a two-domain decoder is good, while the model only using the two-domain codebook (Row 3 of Table 5) and that using the two-domain encoder and codebook (Row 6 of Table 5) demonstrate the worst effect. In short, these conclusions are consistent with the conclusions obtained in the previous section.

Table 5 Experimental results of user study (%)

Model	Ink wash painting			Colored painting		
	Authenticity	Semantic correlation	Average	Authenticity	Semantic correlation	Average
Original VQGAN model	20.34	17.97	19.15	19.66	19.32	19.49
Two-domain encoder	7.12	10.84	9.98	4.40	7.45	8.12
Two-domain codebook	4.06	5.08	4.57	3.72	4.73	4.22
Two-domain decoder	13.22	13.90	13.56	14.57	14.23	14.40
Two-domain encoder and decoder	12.20	10.84	11.52	14.23	11.18	12.71
Two-domain encoder and codebook	4.41	8.14	6.27	3.38	7.45	5.41
Two-domain codebook and decoder	12.20	10.17	11.19	11.52	10.16	10.84
Complete model	**26.44**	**23.05**	**24.75**	**28.47**	**25.42**	**26.95**

5.5 Qualitative analysis

The generated ink wash paintings and colored paintings by the multi-domain VQGAN combined with WenLan 2.0 (1,000 iterations) were given to intuitively present the effect of the proposed model, and the ink wash paintings generated by the CogView (online demo) were also presented. The comparison results are shown in Figures 9, 10, and 11. It can be seen that the results generated by the CogView are relatively simple and are inferior to the results of the proposed model in terms of color, object richness, description details, and semantic relevance. All in all, the generation results of the proposed model have a stronger visual impact and better semantic consistency between text and images.

The method proposed in this paper used a Tesla V100-PCIE-32 GB GPU, and the number of iterations was set to 1,000. The resolution of generated images was 256×256, and the average time to generate each image was about 213.206 s. Under the same hardware conditions, the method described in DALL-E[5] was reproduced, and the model parameters were also set as described in the paper. In addition, it took 569.482 s on average to generate an image of 256×256. It can be seen that the model proposed in this paper is also better in generation efficiency.

6 Summary

Using text descriptions to generate images is a highly challenging task. This paper proposed a text-to-Chinese painting method based on multi-domain VQGAN. The current text-to-image tasks are mostly based on English text, and the generated images have limited domains; the generative method requires massive labeled image–text pairs for training. Considering these problems, this paper proposed to generate multi-domain images with multi-domain VQGAN and used text to guide the generation process of multi-domain VQGAN combined with WenLan. It makes it possible to use Chinese poems to generate multi-domain Chinese paintings. This study performed the ablation experiments, users study, and the comparison of the images generated by this method with the images generated by CogView. It is fully verified that the method proposed in this paper can better generate high-resolution Chinese paintings according to the text semantics.

Text	Multi-domain VQGAN: Ink wash painting	Multi-domain VQGAN: Colored painting	CogView: Ink wash painting
The sun beyond the mountain bows. The Yellow River seawards flows.			
Clouds are worth nothing except those over the mount of Wu.			
Snowy and cold mountains.			
The Yellow River rises as high as white clouds.			
Mountain peaks stand silent as I linger in the slanting sun.			

Figure 9 Comparison of generation results of multi-domain VQGAN and CogView (I)

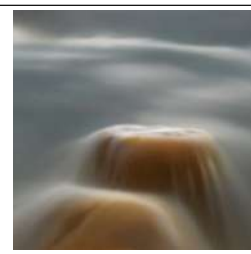
Text	Multi-domain VQGAN: Ink wash painting	Multi-domain VQGAN: Colored painting	CogView: Ink wash painting
The Yellow River comes from the sky, rushing into the sea and never come back.			
With monkey's sad adieus the riverbanks are loud; my skiff has left ten thousand mountains far away.			
At the eastern city gate of Luntai I shall see you off. The road ahead along Tianshan Mountains is heavy with snow.			
The end of the open grassland is the Mountains of spring.			
Far off hangs cataract, a stream upended.			

Figure 10 Comparison of generation results of multi-domain VQGAN and CogView (II)

Text	Multi-domain VQGAN: Ink wash painting	Multi-domain VQGAN: Colored painting	CogView: Ink wash painting
Attics in the clouds.			
The waters quietly ebb and flow. The island mountains skyward go.			
Leaving at dawn the White Emperor crowned with cloud, I've sailed a thousand li through canyons in a day.			
The tide has leveled the sand road, The green mountains in the distance are endless.			

Figure 11 Comparison of generation results of multi-domain VQGAN and CogView (III)

References

[1] Kosslyn SM, Ganis G, Thompson WL. Neural foundations of imagery. Nature Reviews Neuroscience, 2001, 2(9): 635–642. [doi: 10.1038/35090055]

[2] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5907–5915. [doi: 10.1109/ICCV.2017.629]

[3] Chen ZL, Wang C, Wu HM, Shang K, Wang J. DMGAN: Discriminative metric-based generative adversarial networks. Knowledge-based Systems, 2020, 192: 105370. [doi: 10.1016/j.knosys.2019.105370]

[4] Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324. [doi: 10.1109/CVPR.2018.00143]

[5] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-shot text-to-image generation. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8821–8831.

[6] Ding M, Yang ZY, Hong WY, Zheng WD, Zhou C, Yin D, Lin JY, Zou X, Shao Z, Yang HX, Tang J. CogView: Mastering text-to-image generation via transformers. Proc. of the 34th Advances in Neural Information Processing Systems. Curran Associates Inc., 2021. 19822–19835.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN. Attention is all you need. Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5998–6008.

[8] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6306–6315.

[9] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis. Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12873–12883. [doi: 10.1109/CVPR46437.2021.01268]

[10] Huo YQ, Zhang ML, Liu GZ, et al. WenLan: Bridging vision and language by large-scale multi-modal pre-training. arXiv:2103.06561, 2021.

[11] Fei NY, Lu ZW, Gao YZ, Yang GX, Huo YQ, Wen JY, Lu HY, Song RH, Gao X, Xiang T, Sun H, Wen JR. WenLan 2.0: Make AI imagine via a multimodal foundation model. arXiv:2110.14378, 2021.

[12] Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D. DRAW: A recurrent neural network for image generation. Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: PMLR, 2015. 1462–1471.

[13] Wu H, Xu D. Survey of digital image compositing. Journal of Image and Graphics, 2012, 17(11): 1333–1346. [doi: 10.11834/jig.20121101]

[14] Reed SE, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. Proc. of the 33rd Int'l Conf. on Machine Learning. New York City: PMLR, 2016. 1060–1069.

[15] Reed S, Akata Z, Mohan S, Tenka S, Schiele B, Lee H. Learning what and where to draw. Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 217–225.

[16] Hu T, Li JL. Text to image generation based on single-stage GANs. Information Technology and Network Security, 2021, 40(6): 50–55. [doi: 10.19358/j.issn.2096-5133.2021.06.009]

[17] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas DN. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947–1962. [doi: 10.1109/TPAMI.2018.2856256]

[18] Bodla N, Hua G, Chellappa R. Semi-supervised FusedGAN for conditional image generation. Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 689–704. [doi: 10.1007/978-3-030-01228-1_41]

[19] Zhang ZZ, Xie YP, Yang L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt

Lake City: IEEE, 2018. 6199 –6208. [doi: 10.1109/CVPR.2018.00649]

[20] Gao LL, Chen DY, Song JK, Xu X, Zhang DX, Shen HT. Perceptual pyramid adversarial networks for text-to-image synthesis. Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 1019. [doi: 10.1609/aaai.v33i01.33018312]

[21] Lai WS, Huang JB, Ahuja N, Yang MH. Deep Laplacian pyramid networks for fast and accurate super-resolution. Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 624–632. [doi: 10.1109/CVPR.2017.618]

[22] Qiao TT, Zhang J, Xu DQ, Tao DC. MirrorGAN: Learning text-to-image generation by redescription. Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1505–1514. [doi: 10.1109/CVPR.2019.00160]

[23] Tan XY, He XH, Wang ZY, Luo XD, Qing LB. Text-to-image generation technology based on Transformer cross attention. Computer Science, 2022, 49(2): 107–115 (in Chinese with English abstract). [doi: 10.11896/jsjkx.210600085]

[24] Creswell A, Bharath AA. Inverting the generator of a generative adversarial network. IEEE Trans. on Neural Networks and Learning Systems, 2019, 30(7): 1967–1974. [doi: 10.1109/TNNLS.2018.2875194]

[25] Abdal R, Qin YP, Wonka P. Image2StyleGAN: How to embed images into the StyleGAN latent space? Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4432–4441. [doi: 10.1109/ICCV.2019.00453]

[26] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4401–4410. [doi: 10.1109/CVPR.2019.00453]

[27] Abdal R, Qin YP, Wonka P. Image2StyleGAN++: How to edit the embedded images? Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8296–8305. [doi: 10.1109/CVPR42600.2020.00832]

[28] Voynov A, Babenko A. Unsupervised discovery of interpretable directions in the GAN latent space. Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 9786–9796.

[29] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[30] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical Programming, 1989, 45(1): 503–528. [doi: 10.1007/BF01589116]

[31] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 2001, 9(2): 159–195. [doi: 10.1162/106365601750190398]

[32] Zhu JY, Krähenbühl P, Shechtman E, Efros AA. Generative visual manipulation on the natural image manifold. Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 597–613. [doi: 10.1007/978-3-319-46454-1_36]

[33] Huh M, Zhang R, Zhu JY, Paris S, Hertzmann A. Transforming and projecting images into class-conditional generative networks. Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 17–34. [doi: 10.1007/978-3-030-58536-5_2]

[34] Guan SY, Tai Y, Ni BB, Zhu FD, Huang FY, Yang XK. Collaborative learning for faster StyleGAN embedding. arXiv:2007.01758, 2020.

[35] Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114, 2013.

[36] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2672–2680.

[37] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: UNiversal image-TExt representation learning. Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 104–120. [doi: 10.1007/978-3-030-58577-8_7]

[38] Li XJ, Yin X, Li CY, Zhang PC, Hu XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi Y, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[39] Lin JY, Men R, Yang A, Zhou C, Ding M, Zhang YC, Wang P, Wang A, Jiang L, Jia XY, Zhang J, Zhang JW, Zou X, Li ZK, Deng XD, Liu J, Xue JB, Zhou HL, Ma JX, Yu J, Li Y, Lin W, Zhou JR, Tang J, Yang HX. M6: A Chinese multimodal pretrainer. arXiv:2103.00823, 2021.

[40] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[41] Li G, Duan N, Fang YJ, Gong M, Jiang DX. Unicoder-Vl: A universal encoder for vision and language by cross-modal pre-training. Proc. of the 2020 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 11336–11344. [doi: 10.1609/aaai.v34i07.6795]

[42] Su WJ, Zhu XZ, Cao Y, Li B, Lu LW, Wei FR, Dai JF. VL-BERT: Pre-training of generic visual-linguistic representations. arXiv:1908.08530, 2019.

[43] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[44] Sun SQ, Chen YC, Li LJ, Wang SH, Fang YW, Liu JJ. LightningDOT: Pre-training visual-semantic embeddings for real-time image-text retrieval. Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021. 982–997.

[45] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.

[46] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[47] Fu FF. Neural network methods for multi-style Chinese art paintings generation [MS. Thesis]. Chengdu: Sichuan University, 2021. [doi: 10.27342/d.cnki.gscdu.2021.000359]

[48] Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: Diverse image synthesis for multiple domains. Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8188–8197. [doi: 10.1109/CVPR42600.2020.00821]

[49] Tan MX, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 6105–6114.

[50] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

[51] van den Oord A, Li YZ, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018.

[52] Xue A. End-to-end Chinese landscape painting creation using generative adversarial networks. Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3863–3871. [doi: 10.1109/WACV48630.2021.00391]

[53] 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. Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6626–6637.

[54] Szegedy C, Vanhoucke V, Ioffe S, Shles J, Wojna Z. Rethinking the inception architecture for computer vision. Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2818–2826. [doi: 10.1109/CVPR.2016.308]

[55] Frolov S, Hinz T, Raue F, Hees J, Dengel A. Adversarial text-to-image synthesis: A review. Neural Networks, 2021, 144: 187–209. [doi: 10.1016/j.neunet.2021.07.019]

Zelong Sun, master's degree candidate. His research interests include machine learning and text-to-image generation.

Nanyi Fei, Ph.D. candidate. His research interests include computer vision and visual-language-oriented multimode.

Guoxing Yang, Ph.D. candidate. His research interests include machine learning and image generation.

Zhiwu Lu, Ph.D., professor, doctoral supervisor. His research interests include machine learning and computer vision.

Jingyuan Wen, master's degree candidate. His research interests include multimodal learning and computer vision.

Jirong Wen, Ph.D., professor, doctoral supervisor. His research interests include information retrieval, data mining, machine learning, and databases.