Learning to Generate Poetic Chinese Landscape Painting with Calligraphy

Shaozu Yuan[1] , **Aijun Dai**[1] , **Zhiling Yan**[1] , **Ruixue Liu**[1] , **Meng Chen**[1] ,
Baoyang Chen[2] , **Zhijie Qiu**[2] , **Xiaodong He**[1]
[1]JD AI, Beijing, China
[2]Central Academy of Fine Arts, Beijing, China
{yuanshaozu, daiaijun1, yanzhiling1, liuruixue, chenmeng20, xiaodong.he}@jd.com
{chenbaoyang, qiuzhijie}@cafa.edu.cn

Abstract

In this paper, we present a novel system (denoted as Polaca) to generate poetic Chinese landscape painting with calligraphy. Unlike previous single image-to-image painting generation, Polaca takes the classic poetry as input and outputs the artistic landscape painting image with the corresponding calligraphy. It is equipped with three different modules to complete the whole piece of landscape painting artwork: the first one is a text-to-image module to generate landscape painting image, the second one is an image-to-image module to generate stylistic calligraphy image, and the third one is an image fusion module to fuse the two images into a whole piece of aesthetic artwork[1].

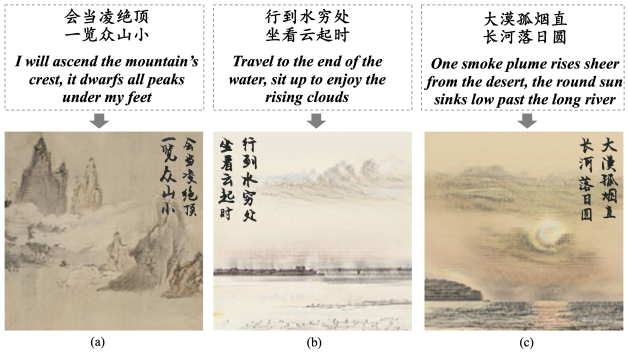

Figure 1: The generated artwork examples of our system.

1 Introduction

Chinese landscape painting, or shan shui ("mountain-water"), is an essential style of traditional Chinese painting which involves or depicts natural landscapes, using a brush and ink rather than conventional paints. Mountains, rivers and waterfalls are common subjects of shan shui paintings. Besides, calligraphy and poetry are usually inseparable from Chinese landscape painting. All three together are regarded as the purest forms of art [Murck *et al.*, 1991]. Poetry expresses the thinking and opinion of the artist. Landscape painting constructs the artistic conception. Calligraphy reflects the emotion and inner psychology of the artist. They complement each other and elevate the artistry of artwork significantly.

With the success of deep learning, Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2014] have been widely applied in artwork generation and made remarkable progress [Xu *et al.*, 2018; Tomei *et al.*, 2019; Ding *et al.*, 2021]. However, to the best of our knowledge, previous works only focus on a single form of artwork, such as poetry generation [Wang *et al.*, 2016; Guo *et al.*, 2019; Shen *et al.*, 2020], calligraphy generation [Zhang *et al.*, 2018; Gao and Wu, 2020; Liu *et al.*, 2020], or Chinese landscape painting generation [Lin *et al.*, 2018; He *et al.*, 2018; Zhou *et al.*, 2019], and no one explores the composition of poetic Chinese landscape painting with calligraphy as a whole piece of artwork. Meanwhile, most of the previous approaches [Lin *et al.*, 2018; He *et al.*, 2018; Zhou *et al.*, 2019] treat the Chinese landscape painting generation as a style transfer problem based on image-to-image translation, which heavily relies on the conditional inputs (e.g. photograph or sketches), thus it is restricted in the number of generated images. Since each of its generation is built upon a single, human-fed input, the traditional style transfer methods tend to produce derivative artworks that are stylistic copies of conditional inputs, which lacks creativity and imagination for the artwork creation.

In this paper, we propose a novel and challenging task of generating poetic Chinese landscape painting with calligraphy. Unlike previous image-to-image approaches, we formulate the task as a text-to-image task and develop a system that takes the classic poetry text as its input and outputs the artistic landscape painting image with calligraphy. Specifically, the generation process is as follows: 1) Landscape painting generation takes the poetry text as input and applies the text-to-image model to generate a poetic painting image that contains the main subjects of the poetry. 2) Calligraphy generation leverages the style transfer model to generate stylistic calligraphy images from standard font images. 3) The image fusion module predicts the layout of calligraphy and fuses the calligraphy image into the landscape painting image. As there is no off-the-shelf poetry-to-painting dataset, we also construct a dataset with more than 5,000 text-image pairs via an automatic method which will be further introduced later.

[1]Demo video: https://youtu.be/xRo8xiTXb74

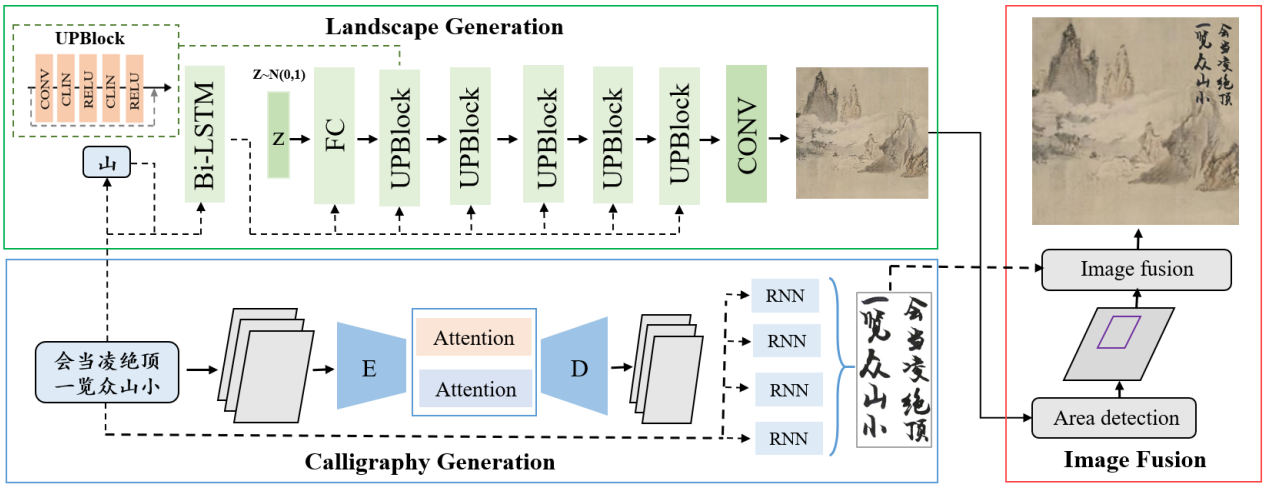

Figure 2: The architecture of our system is composed of three parts: landscape generation, calligraphy generation and image fusion. The input to the system is a classic poetry text, and the output is a landscape painting with calligraphy.

2 Dataset Construction

We collect some Chinese landscape painting images from previous works [Bao *et al.*, 2010; Xue, 2021]. To enrich the data diversity, we also crawl thousands of real landscape photographs from the web. To construct semantic-related poetry-painting pairs, we set up a dataset construction pipeline as Figure 3 shows. We first train a CycleGAN model [Zhu *et al.*, 2017] to transfer the landscape photographs to photographs paintings. To match the classic poetry for each image automatically, we first detect the objects like mountains, rivers and trees in the photograph, then extract key entities (e.g. mountain, water, forest) from poetry via TextRank [Mihalcea and Tarau, 2004], finally match the photograph with poetry based on the overlap of entities and objects. Considering the literal differences between ancient Chinese and the output labels of object detection model, we also construct a mapping dictionary based on synonyms to facilitate the matching. By this way, more than 5,000 text-image pairs are constructed for text-to-image model training.

For the calligraphy generation task, we use the same dataset as [Liu *et al.*, 2020]. We also annotate the location information of calligraphy in 500 traditional landscape paintings to facilitate the training of our image fusion model.

3 System Architecture

As shown in Figure 2, our system Polaca contains three modules: 1) a Chinese landscape painting generation module that creates painting images from the input text, 2) a calligraphy generation module that generates calligraphy images based on the input text, 3) a multi-modality image fusion module which combines the generated calligraphy and landscape painting images with area detection and image fusion.

3.1 Landscape Painting Generation

To guide the landscape painting generation effectively, we design a novel cross-modal generative adversarial network that is composed of a generator, a discriminator, and a text

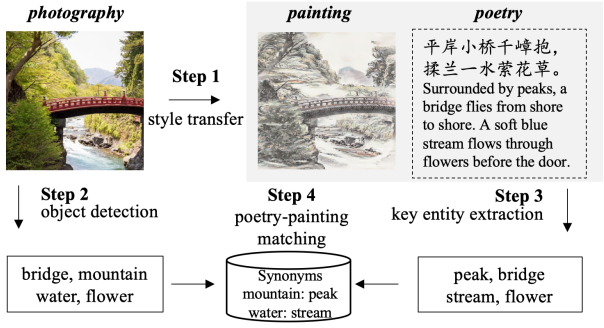

Figure 3: Illustration of dataset construction.

encoder. The text encoder first encodes the poetry and the extracted basic painting elements into a text vector with Bi-LSTM [Huang *et al.*, 2015]. Then, to ensure the diversity of generated images, we feed the initial image vector, sampled from the Gaussian distribution, into the generator. The text and image features are fused by stacked up blocks during the image generation process. In particular, we adopt CLIN (Conditional Layer-Instance Normalization) [Yuan *et al.*, 2021] in up blocks, which combines the advantages of instance normalization and layer normalization to selectively change or keep the content information. Finally, a discriminator is applied to distinguish real images from synthetic images [Goodfellow *et al.*, 2014].

3.2 Calligraphy Generation

Calligraphy generation consists of two sub-modules: character image generation and layout prediction. The character image generation module generates a stylistic calligraphy image from standard font images for every character in the input text. To guarantee the quality of the generated character image, we implement a modified GAN model equipped with two auxiliary classifiers with the attention mechanism

in the generator to catch both the content-aware representation and style-aware representation by following [Kim *et al.*, 2019]. The layout prediction module predicts the overall spatial arrangement for the whole piece of calligraphy artwork. We devise a recurrent neural network [Chung *et al.*, 2014] to solve the sequential modelling problem.

3.3 Image Fusion Module

To fuse the calligraphy image into the landscape image naturally, it is necessary to detect the proper area in the landscape painting image. Here, we devise a layout network based on Faster R-CNN [Ren *et al.*, 2015] to predict the location and size of the calligraphy image. Then we synthesize the final artwork by implementing pixel-level image fusion. With the advantage of Faster R-CNN and pixel operation, the image fusion module can insert the generated calligraphy image into the generated landscape painting appropriately.

4 Evaluation and Analysis

4.1 Human Evaluation

We conduct human evaluation to measure the quality of generated landscape paintings with calligraphy. We first sample 100 real landscape painting images created by human from our dataset, then mix it with 100 generated paintings together to form the test set. Then we ask the human evaluators to score each painting by hiding the image source. Twenty students majoring in fine arts are guided to evaluate the images based on the criteria of landscape quality (landscape), calligraphy quality (calligraphy), coherence between landscape images and poetry (coherence), and aesthetics of the overall artwork (aesthetics). The rating score ranges from 1 to 10 where 10 represents the best quality. Table 1 demonstrates that even though the generated images are rated lower than real artworks in the quality of *landscape* and *calligraphy*, they obtain comparative scores with human paintings in the aspect of *coherence*, indicating our proposed architecture can generate semantic-related landscape paintings by understanding the poetry content.

Scores	Landscape	Calligraphy	Coherence	Aesthetics
Generated	8.55	8.83	7.12	8.08
Human	9.23	9.15	7.22	9.12

Table 1: The user study on generated and real paintings.

4.2 Case Study

Besides, we also make further case study to discuss the advantage and limitation of our system. As shown in Figure 4 (a), our system captured the key elements of *river, creek, mountain* and depicted them in the generated paintings accurately. And the calligraphy is also visually pleasing. We conjecture that all three elements appear frequently in traditional Chinese paintings and poetry which are easy to learn. Meanwhile, it indicates the effectiveness of the CLIN block on catching the interactions of text and image. However, in Figure 4 (b), our model omitted the *moon* and *zither* by mistake. What's more, the hues are also too dark, which damaged the artistic conception of whole painting. In the future, we plan to design more objectives to encourage the model to improve the completeness during text-to-image generation.

Poetry: 江流天地外，山色有无中

Trans: Beyond the horizon, waters of rivers and creeks gush and flow. The mountainous landscape undulates varying scenes.

(a)

Poetry: 松风吹解带，山月照弹琴

Trans: Wind blows through pines to my untied robe. The moon shines on me playing zither at the mountain.

(b)

Figure 4: Case study examples.

Figure 5: The interface of our Demo.

5 Demonstration

In order to facilitate a user-friendly experience, we develop a website for users to experience the journey of art creation. Users can create poetic Chinese landscape paintings with calligraphy by entering Chinese poetry. The system is implemented with TensorFlow and Python, and 4 GPUs (NVIDIA Tesla P40) are deployed for real-time inference. Please watch the demonstration video for more details. Figure 5 shows the demo interface.

6 Conclusion

In this paper, we develop a novel system to challenge the task of composing poetic Chinese landscape painting with calligraphy. The system consists of a text-to-image module to generate landscape painting, an image-to-image module to create stylistic calligraphy images, and an image fusion model to combine the calligraphy and landscape painting images together naturally. We also contribute a large-scale poetry-to-painting multi-modal dataset. A web-based demo is established to make the system easily accessible.

References

[Bao *et al.*, 2010] Hong Bao, Ye Liang, Hong-zhe Liu, and De Xu. A dataset for research of traditional chinese painting and calligraphy images. In *Proceedings of the 2nd International Conference on Information and Multimedia Technology. Singapore: IACSIT Press*, pages 136–143, 2010.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[Ding *et al.*, 2021] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021.

[Gao and Wu, 2020] Yiming Gao and Jiangqin Wu. Gan-based unpaired chinese character image translation via skeleton transformation and stroke rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 646–653, 2020.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Guo *et al.*, 2019] Xiaoyu Guo, Meng Chen, Yang Song, Xiaodong He, and Bowen Zhou. Automated thematic and emotional modern chinese poetry composition. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 433–446. Springer, 2019.

[He *et al.*, 2018] Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1172–1180, 2018.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Kim *et al.*, 2019] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[Lin *et al.*, 2018] Daoyu Lin, Yang Wang, Guangluan Xu, Jun Li, and Kun Fu. Transform a simple sketch to a chinese painting by a multiscale deep neural network. *Algorithms*, 11(1):4, 2018.

[Liu *et al.*, 2020] Ruixue Liu, Shaozu Yuan, Meng Chen, Baoyang Chen, Zhijie Qiu, and Xiaodong He. Maliang: An emotion-driven chinese calligraphy artwork composition system. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4394–4396, 2020.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[Murck *et al.*, 1991] Alfreda Murck, Wen C Fong, and Wen Fong. *Words and images: Chinese poetry, calligraphy, and painting*. Metropolitan Museum of Art, 1991.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[Shen *et al.*, 2020] Lei Shen, Xiaoyu Guo, and Meng Chen. Compose like humans: Jointly improving the coherence and novelty for modern chinese poetry generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Tomei *et al.*, 2019] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019.

[Wang *et al.*, 2016] Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*, 2016.

[Xu *et al.*, 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[Xue, 2021] Alice Xue. End-to-end chinese landscape painting creation using generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3863–3871, 2021.

[Yuan *et al.*, 2021] Shaozu Yuan, Ruixue Liu, Meng Chen, Baoyang Chen, Zhijie Qiu, and Xiaodong He. Learning to compose stylistic calligraphy artwork with emotions. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3701–3709, 2021.

[Zhang *et al.*, 2018] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8447–8455, 2018.

[Zhou *et al.*, 2019] Le Zhou, Qiu-Feng Wang, Kaizhu Huang, and Cheng-Hung Lo. An interactive and generative approach for chinese shanshui painting document. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 819–824. IEEE, 2019.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.