Multimodal Fusion for Traditional Chinese Painting Generation

Sanbi Luo, Si Liu, Jizhong Han, and Tao Guo(✉)

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
{luosanbi,liusi,hanjizhong,guotao}@iie.ac.cn

Abstract. Creativity is a fundamental feature of human intelligence, and a challenge for artificial intelligence (AI). In recent years, AI has gained tremendous development in many single tasks with single models, such as classification, detection and parsing. As the development continued, AI has been increasingly used for more complex tasks, multitasking for example, and then research in multimodal fusion naturally became a new trend. In this paper, we propose a multimodal fusion framework and system to generate traditional Chinese paintings. We select suitable existing networks for different elements generation in this oldest continuous artistic traditions artwork, and finally fusion these networks and elements to create a complete new painting. Meanwhile, we propose a divide-and-conquer strategy to generate large images with limited GPU resources. In our end-to-end system, a large image becomes a traditional Chinese painting in minutes automatically. It shows that our multimodal fusion framework works well and AI methods has good performance in traditional Chinese painting creation.

Keywords: Multimodal fusion · Image style · GAN
Traditional Chinese painting

1 Introduction

Boden said in 1998 [1], "Creativity is a fundamental feature of human intelligence, and a challenge for AI." Nearly 20 years later, AI has made astonishing development, especially in recent years, the big bang of AI has been fueled by deep learning (DL). Neural networks, such as RNN and GAN, can generate poetry and images. In this paper, we plan to use DL to generate traditional Chinese paintings, which is a multi-tasking challenge that requires the fusion of many DL networks and technologies.

Traditional Chinese painting is one of the oldest continuous artistic traditions in the world. As show in Fig. 1, it usually consists of four parts: painting, blank space, poem and calligraphy, and seal [3]. Painting is the most important and essential part in traditional Chinese painting. It has a unique style, which emphasis on spirit rather than realism. Blank space is a philosophical concept in traditional Chinese painting. It is given a lot of meaning. Here we only consider it as a form of composition. There are usually poems and seals in it. The poems are usually an expression and a supplement of a traditional Chinese painting. Calligraphy is understood in China as Chinese art of

© Springer Nature Switzerland AG 2018
R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 24–34, 2018.
https://doi.org/10.1007/978-3-030-00764-5_3

writing a good hand with the brush or the study of the rules and techniques of this art. Seals always represent the author, or owners.

To generate a traditional Chinese painting means at least has to face following challenges:

- Style transfer a painting from a image
- Find and generate the blank space (optional)
- Generate poems about the painting
- Generate the seal of the author
- Transfer poems and seals to a Chinese calligraphy style
- Fusion all elements to create a traditonal Chinese painting

In this paper, we chose the latest and most suitable networks, such as YOLO, Mask RCNN, Neural Style Transfer network, Cycle GAN, pix2pixHD and LSTM, to be part of our fusion system. We select the key parts of these networks and integrate them into a fusion one and finally used for generating traditional Chinese paintings.

Fig. 1. A traditional Chinese painting. It always has four elements: painting, blank space, poem and calligraphy, and seal

2 Related Work

Many works can generate images. They can be roughly divided into three categories: GANs [4] and GANs-based networks [5–8], RNN-bested networks [9–11], and some other methods [12–14, 16, 17]. These methods can generate hand-written numbers, human faces, indoor and outdoor scenes, and something else.

There are also many works that can generate stylized-images. Cycle-GAN [8] is an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples, such as to translate a horse to a zebra. In [12], authors use neural representations to separate and recombine content and style of arbitrary images, providing a neural algorithm for the creation of artistic images, which is based on VGG [15] network. However, these networks are always limited to GPU's

performance and cannot translate large images directly with limited GPU resources. With a 12G Video Random Access Memory (VRAM) GPU, these methods always hard to generate more than 1024 * 1024 pixel size images. To generate larger images, more GPU VRAM is needed. If one wants to translate a big image as 10240 * 10240 pixel size, 1200G VRAM GPU may be necessary. GPUs are always very expensive, and it is often a bottleneck in image processing.

Scene parsing, or recognizing and segmenting objects in an image, is one of the key problems in computer vision. It can recognize and segment objects in an image, such as the sky and mountains. It also can be used in identifying and segmenting blank space and painting in traditional Chinese paintings. A great deal of works has gained remarkable achievements, such as FCN [22] and DeepLab [2]. ADE20K [18] collects precise dense annotation of scenes, objects, parts of objects with a large and open vocabulary. It works well in image parsing and provides a good code implementation in GitHub.

Object detection is a core problem in computer vision too. There are many methods available, such as Deformable parts models (DPM), a sliding window approach to object detection [25], or R-CNN and its variants [24], using region proposals instead of sliding windows to find objects in images. In [19], authors introduced YOLO, a real-time detection system. This model is simple to construct and can be trained directly on full images, and it is one of the fastest object detector at present.

The success of recurrent neural network (RNN) models in complex tasks like machine translation and audio synthesis has inspired immense interest in learning from sequence data. In [26], the author proposed a character-level Language model to generate sentences in English. In [20], the author trained an RNN character-level language model on Chinese poetry datasets, and it shows that RNN-based models can generate Chinese poetry well.

Multimodal fusion learning has been extensively studied in recent years, such as Visual Question Answering (VQA) [27]. As single neural models for single tasks become efficient and practical, Multimodal fusion naturally receives attention and development when AI faces more complex challenges.

3 Traditional Chinese Painting Generation

Figure 2 shows the framework in our system how to generate traditional Chinese paintings by fusion network. With the input of a content image and a style image, our system generates four elements through four branches, and finally merges all branches to generate a traditional Chinese painting. Branch 1 is used to generate poetry and calligraphic styles. It usually requires CNN, RNN and GAN networks. Branch 2 is optional. It is used to generate blank space in traditional Chinese paintings. We can use parsing networks to erase some non-critical parts in picture, such as the sky, and in traditional Chinese painting the sky is always blank. Branch 3 is used to generate stylized large paintings. The famous Neural Style Transfer network and GAN nets could be used for reference. Branch 4 is used to make seals, and a stylized network is required.

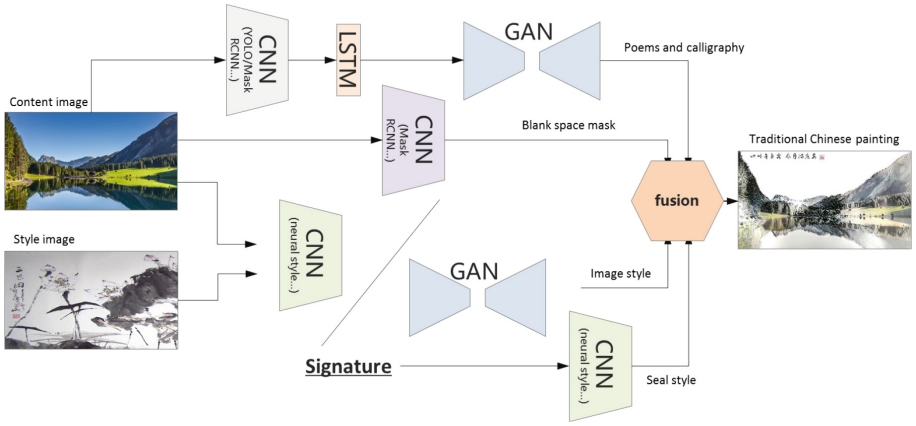

Fig. 2. The multimodal fusion framework of our system in traditional Chinese painting generation

The characteristic of our fusion framework is that it focuses on the selection and integration of sub-networks. Therefore, on the one hand we can choose and integrate the latest and best single network into our framework to generate traditional Chinese painting elements, on the other hand we must find a balance between elements fusion and networks fusion, and it brings us challenges and innovation opportunities in fusion research. Just like human brains integrate a variety of information to make a decision, fusion will be one of the trends in the future of AI industry. In this version of our system, we used the key parts of existed YOLO, Mask RCNN, Neural Style Transfer network, Cycle GAN and LSTM-based network to form the multimodal fusion framework. We have integrated all the networks in branch 1and branch 4, and we fuse all the elements generated by 4 branches to draw a painting.

3.1 A Divide-and-Conquer Strategy for Large Image Style Transfer

In branch 3, the original Neural Style Transfer Nets is hard to generate a picture larger than 1024 * 1024 pixel size with 12G GPU. Pix2pixHD [23] can generate 1024 * 2048 pixel size pictures with 12G GPU, but it requires thousands of big paintings for training, and so many big paintings are very hard to collect. Therefore, we propose a divide-and-conquer strategy for large image style transfer, which can generate large paintings with a single content and a single style image. There are three steps to generate style paintings: divide, stylize, and merge.

Step 1: split larger images into smaller ones. When dividing an original large content image into sub-content-images, we let the adjacent sub-content-images have overlaps. These overlaps will be used in step 2 to guide the stylization of adjacent sub-content-images.

As shown in Fig. 3, if we divide an image like Fig. 3(A) without overlaps, the stylized image merged by sub-stylized-images will probably like Fig. 3(E). It does not look like a single image but a puzzle with four pieces. Therefore, we try to divide the

image like Fig. 3(B), to make every two horizontally and vertically adjacent sub-content-images have an overlap, and then control another sub-content-image's stylization to make the stylized overlaps look the same with the former sub-stylized-image in step 2. The red masks in Fig. 3(C) and (D) are overlaps between two sub-content-images, and the overlap is half of each sub-content-image in this case. In our experiments, we found that either L1loss (in pytorch) of overlap is less than 3e−2 or MSELoss (in pytorch) less than 1.3e−3, it always cannot be distinguished one sub-stylized-image from the other parts with naked eyes. In Fig. 3(F), L1loss is 7.0461e−02 and MSELoss is 7.1458e−03, we can distinguish between the left and right parts. In Fig. 3(G), L1loss is 3.0060e−02 and MSELoss is 1.3479e−03, the left and right parts look like one picture with naked eyes. In Fig. 3(H), L1loss is 9.9688e−03 and MSELoss is 1.7576e−04, and it looks like a single image too. So it has to make the L1losses of overlap between two stylized adjacent sub-images less than 3e−2, or MSELoss less than 1.3e−3. We call the loss between the overlaps of two adjacent images similar1 loss.

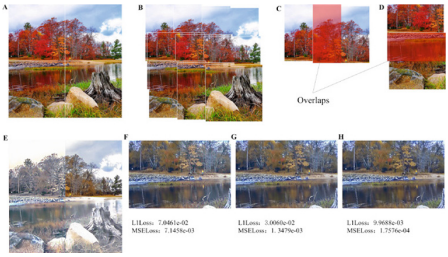

Fig. 3. Image division with overlap. (Color figure online)

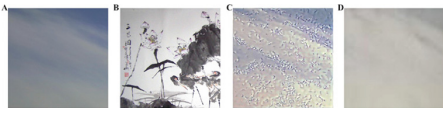

Fig. 4. A problem in sub-image transfer. A is a sub-content-image, B is style image, C is a sub-stylized-image that the problem has occurred, and D is the image that a sub-stylized-image should be.

Fig. 5. Sub-content-images stylization one by one.

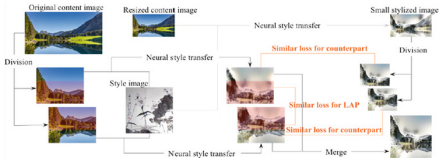

Fig. 6. Two similar losses for image transfer.

Step 2: transfer sub-content-images one by one. Our strategy is that at first a large image is divided into sub-content-images, and then they are translated sub-content-images in a certain order gradually. As shown in Fig. 5, we firstly translate sub-content-image 1 to a sub-stylized-image according to the style image. Then translate sub-content-image 2 depends on the style image and sub-stylized-image 1, and so on until the end of the first line. Sub-content-image 6 should depend on the style image and sub-stylized-image 1. Sub-content-image 7 has to depend on the style image, sub-stylized-image 6 and sub-stylized-image 2. It seems like the generate order of

PixelRNN [16]. The difference is that PixelRNN is for pixels, and our strategy is for sub-content-images.

As mentioned in step 1, if we divide an image into sub-content-images using Neural Style Transfer network [12] without overlaps, the merging stylized image may look like a puzzle with many parts. Therefore we use divide-and-conquer strategy with overlaps to deal with such a problem. Also, there is another problem that if the sub-content-image is very monotonous, a whole piece of blue sky or calm water for instance, the sub-stylized-images will not be monotonous but tend to the content of the style image. It is determined by the structure of Neural Style Transfer networks [12]. As shown in Fig. 4, Fig. 4(A) is a sub-content-image, Fig. 4(B) is style image, Fig. 4 (C) is a sub-stylized-image that the problem has happened, and Fig. 4(D) is the image that a sub-stylized-image should be. To solve this problem, we added a new loss between the sub-stylizing-image and the corresponding area of the stylized image, which is a small one resized from the original large content image. We call this loss as similar2 loss.

As illustrated in Fig. 6, we use two losses to control sub-content-image stylization. The one is to improve the similarity of overlaps between adjacent sub-stylized-images, and the other is to eliminate the problem caused by local sub-content-image stylization. Resized content image is minified from the original content image, a small one (under 1024 * 1024 pixel size) that can be stylized by a 12G VRAM GPU at a time.

Let p and a be the original content image and the style image, f and c be the overlap of the former stylized image and the corresponding area of the minified and stylized image, and x be the image that is generated. The loss function that should be minimized is:

$$L_{total}(p,a,f,c,x) = \alpha L_{content}(p,x) + \beta L_{style}(a,x) + \gamma L_{similar1}(f,x) + \delta L_{similar2}(c,x) \quad (1)$$

α, β, γ and δ are the weighting factors for content, style, similar1 and similar2 loss respectively.

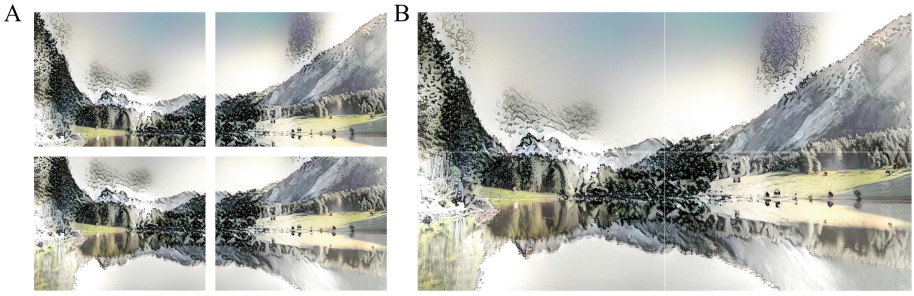

Fig. 7. Generate large stylized image (1920 * 1080 pixel size).

Step 3: merge the stylized small images into a big picture. When sub-content-images style translation is finished, it is time to merge sub-stylized-images to a big one. Figure 7 shows the process of picture integration. The merged styled-image consists of four sub-stylized-images, show in Fig. 7(B), and each part is 960 * 540 pixel size and divided from the 1024 * 800 pixel size sub-stylized-image separately, show in Fig. 7 (A). The experiment uses a 12G VRAM GPU, which can only generate 1024 * 1024 pixel resolution images at most by using Neural Style Transfer Nets directly. It proves that our divide-and-conquer strategy with neural style translate networks could transfer 1920 * 1080 pixel size image by a 12G VRAM GPU, which can only transfer a 1024 * 1024 pixel size image using Neural Style Transfer network directly.

3.2 Find and Generate the Blank Space

In [18], the authors introduces a new densely annotated dataset with the instances of stuff, objects, and parts, covering a diverse set of visual concepts in scenes. A generic network design was proposed to parse scenes into stuff, objects, and object parts in a cascade. We use these image semantic segmentation achievements to find and generate the blank space in stylized images, showing in upper parts of paintings in Fig. 9.

3.3 Generate Poetry and Stylize into Chinese Calligraphy Style

In [19], the author introduces a real-time detection system YOLO. It can recognize the objects in the content image, such as sky, person, mountain, house, water, and so on. Meanwhile, we found that RNN-based poetry generate model [20] is very suitable for our system to generate Chinese poems. We combined these two networks to form a fusion network to generate poetics from pictures, and combined a trained GAN-based neural network [21] to stylize the text. In Fig. 8(A), branch 1 of our framework generated Chinese poems for the picture of the first row and first column in Fig. 9. It means the mountain is moist with water surrounding, and the house and yard are full of romance. It sounds objective and poetic.

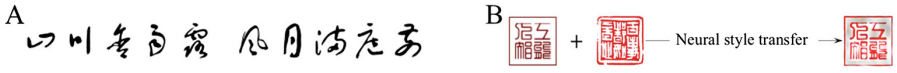

Fig. 8. A Chinese calligraphy style poem and a seal stylization.

3.4 Generate and Stylize the Seal

Seals in traditional Chinese paintings represent the author or owners. They use Small Seal Script. Small Seal Script cannot be transferred from the other fonts for its special structure, so we use TrueType fonts directly. However, you can use Neural Style Transfer network to stylize the seal. As shown in Fig. 8(B), the first seal on the left is generated by TrueType font of Small Seal Script, and it means “artificial intelligence”. The rightmost one is stylized by a real seal image.

3.5 Create a Complete Traditional Chinese Painting

After four elements of traditional Chinese painting have been generated, we can integrate these four elements to generate a large traditional Chinese painting, as shown in Fig. 9.

Fig. 9. Traditional Chinese paintings generated by our system. The first column is landscape photos, and the second column is traditional Chinese paintings created by our system

4 Results and Limitation

Figure 9 shows some results of our system. The first column is landscape photos, and the second column is traditional Chinese paintings created by our system, which contain painting, blank space, poem and calligraphy, and seal. The first row is a

1920 * 1080 pixel size. The second is 3824 * 2144 pixel size. The third and fourth rows are all 2600 * 916 pixel size. In 12G VRAM GPU environment, we divide the content image into suitable size sub-content-images, generally no more than 1024 * 1024 pixel size. It cost 10–20 min. These successful cases with divide-and-conquer strategy show that our system works well.

Table 1 shows that compared to the current mainstream methods, our method can also generate poem, seal and blank space instead of just generating big size stylized images. Although our system could generate larger images than that Neural Style Transfer network can do with limited GPU resources. Due to the divide-and-conquer strategy, the generation speed depends on both the speed of Neural Style Transfer network and the size of the picture. Meanwhile, we use semantic segmentation to generate blank space and use RNN character-level language model and GAN-based networks to generate poem and calligraphy. Naturally the quality of the generated traditional Chinese paintings also depends on the development of these technologies. In the aspect of multimodal fusion, we also operate in elements fusion and networks fusion together. These are the limitations of our system.

Table 1. Comparison of painting generation methods.

Heading level	Image style	Big size	Poem	Seal	Blank space
Neural style [12]	✓				
Cycle GAN [8]	✓				
Pix2pixHD [23]	✓	✓			
Our system	✓	✓	✓	✓	✓

5 Conclusion and Future Work

In this paper, we propose a multimodal fusion framework and system to generate traditional Chinese paintings. We select suitable existing networks for different elements generation in this oldest continuous artistic traditions artwork, and finally fusion these networks and elements to create a new painting. Meanwhile, we propose a divide-and-conquer strategy to generate large images with limited GPU resources. In our end-to-end system, a large image becomes a traditional Chinese painting in minutes automatically. It shows that our solution works effectively and AI methods has good performance in traditional Chinese painting creation. We believe that just like human brains integrate a variety of information to make a decision, fusion network research will be one of the trends in the future of AI industry. Next we will continue to focus on multimodal fusion research.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Nos. U1536203, 61572493), IIE project (No. Y6Z0021102, No. Y7Z0241102) and CCF-Tencent Open Research Fund.

References

1. Boden, M.A.: Creativity and artificial intelligence. Artif. Intell. **103**(1/2), 347–356 (1998)
2. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
3. http://oilpaintingfactory.com/traditional-Chinese-painting.html
4. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
5. Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. In: Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester (2014)
6. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
8. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)
9. Gregor, K., Danihelka, I., Graves, A., Wierstra, D.D.: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)
10. van den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759 (2016)
11. Yang, J., Reed, S., Yang, M.-H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In: NIPS (2015)
12. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
13. Tieleman, T.: Optimizing neural networks that generate images. Ph.D. thesis, University of Toronto (2014)
14. Dosovitskiy, A., Springenberg, J., Tatarchenko, M., Brox, T.: Learning to generate chairs, tables and cars with convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2016). https://doi.org/10.1109/TPAMI.2016.2567384
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
16. van den Oord, A., et al.: Conditional image generation with PixelCNN decoders. CoRR, abs/1606.05328 (2016)
17. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv:1605.05396 (2016)
18. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene paring through ADE20K dataset. In: Proceedings of CVPR (2017)
19. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: CVPR (2017)
20. https://github.com/justdark/pytorch-poetry-gen
21. https://github.com/kaonashi-tyc/zi2zi
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
23. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. arXiv preprint arXiv:1711.11585
24. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: NIPS (2016)

25. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
26. Karpathy, A.: The unreasonable effectiveness of recurrent neural networks. Andrej Karpathy Blog (2015). http://karpathy.github.io
27. Antol, S., et al.: VQA: visual question answering. In: ICCV (2015)