Generating Chinese Ancient Ink Painting from Sketch Using Conditional-GANs

Chenchen Ma
Department of Statistics
University of Michigan
chenchma@umich.edu

Kang Liu
Department of Electrical Engineering
and Computer Science
University of Michigan
kangliu@umich.edu

Zihao Deng
Department of Materials Science
and Engineering
University of Michigan
zhdeng@umich.edu

Xiaofan Zhu
Department of Electrical Engineering
and Computer Science
University of Michigan
zhuxf@umich.edu

Abstract

Generating Chinese ancient ink paintings is challenging due to the blurred and abstract features in the paintings that may not be well-captured by the conventional neural style transfer or GANs. To improve the artistic quality of the generated ink paintings, we incorporate and experiment with different additional losses in the framework of conditional-GAN (cGAN) Pix2Pix, and propose a new cGAN architecture (Ske2Ink) which incorporates category information of the target paintings by the introduction of category embeddings. With this proposed model, we demonstrate good quality ink painting generation scenarios both from Chinese paintings as well as real photos. We find that L2 + style loss is the most suitable additional loss for Chinese ink painting generation task, and our Ske2Ink model performs well on the mix-category dataset.

1 Introduction

Ink painting is one representative form of ancient Chinese art. These paintings often portray objects that are commonly found in nature, such as flowers, mountains and rivers, etc. Different from western paintings, which largely focus on the realistic characteristics of subjects, Chinese ink paintings emphasize more on the abstract aspect of the depicted objects by using sparse brush strokes. As a result, most of the ink paintings tend to appear blurred rather than sharp and clear, which also reflects the way ancient Chinese artists view the nature.

Recently, deep-learning-based generative models have received huge attentions due to their great performance on generating realistic images that are indistinguishable from real images created by humans. It then follows that art works including ancient Chinese ink paintings can also be created by algorithms based on similar methodologies. In fact, ancient Chinese painting generation has been investigated in the framework of neural style transfer [1] as well as generative adversarial networks (GAN). [2, 3, 4] However, the quality of ink paintings is not satisfactory if they are generated directly from neural style transfer or GAN without imposing additional constraints. This is because Chinese art tends to create a hazy atmosphere to show the appreciation of spirit rather than appearance, and it is challenging to reach a balance between clearness and abstractness. Thus, additional terms in the loss function will be necessary to have the generated ink paintings possess the abstract features of Chinese art. Also, multimodal Chinese style transfer or diverse style painting generation has not been achieved very successfully, which is another challenge in Chinese ink painting generation. In this respect, we aim to generate high quality ink paintings with Chinese style from sketch figures, and explore

EECS 598 Deep Learning, Final Report, April 2019

diverse style painting generation via a series of cGAN architectures. An illustration of transferring line sketch to Chinese painting is shown in Figure 1. To achieve this goal, we divide our project into three parts. First, we investigate the influence of different additional losses to the style of the generated paintings based on a baseline cGAN architecture. Then, we generalize this cGAN to incorporate category embeddings for mix-category dataset. Finally, we try to generate Chinese ink paintings with diverse styles.

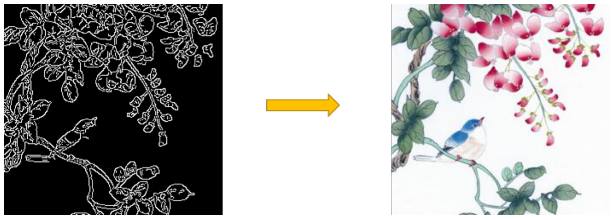

Figure 1: Example image of a line sketch and its original image

We make the following contributions to the community of Chinese painting generation:

- We compare the influence of the additional losses on the style of the generated paintings both qualitatively and quantitatively, and find that L2 + style loss based on Gram matrix is the most suitable one for generating Chinese ink paintings.
- We propose a new cGAN architecture called Ske2Ink built on top of the baseline Pix2Pix model to accommodate the situation where ground truth paintings belong to different categories. Our model has a much better performance than Pix2Pix on a mix-category dataset including flowers and mountains. Our model also generates reasonable paintings based on real photos.
- We investigate the possibility of diverse style ink painting generation by incorporating randomness into the category embedding, and analyze the reason why it fails to produce diverse style paintings.
- We demonstrate that BicycleGAN [5] is not able to achieve diverse-style Chinese ink painting generation with good quality.

2 Related Work

In this section, we briefly review related work on deep-learning-based generative models, and ancient Chinese painting generation.

2.1 Deep-learning-based generative models

The goal of the generative model is to learn the distribution of the model to match the distribution of the training data, either explicitly or implicitly. Van der Oord et al [6] propose pixelCNN/pixelRNN framework to generate the image pixel by pixel, which is slow for sampling even though the model offer stable training. Instead of maximizing the likelihood of the training data using a tractable probability density, Kingma and Welling [7] introduce auto-encoding variational bayes to maximize the lower bound of the posterior distribution. This provides faster training and inference. However, the generated images are often blurred despite easy control through the latent variables. Compared to above mentioned explicit models, learning the distribution implicitly based on generative adversarial network (GAN) [8] can produce sharp and realistic images with high quality. Since then, GAN has been widely applied to various image generation tasks, and researchers have designed different architectures for the generator and discriminator within the GAN framework to improve the performance, e.g. DCGAN [9], WGAN [10], cycleGAN [11], etc. Due to the high performance of GAN on image generation, we employ cGAN to generate Chinese ink paintings in this project.

2.2 Ancient Chinese painting generation

Even before deep learning becomes prevalent, many researchers [12, 13, 14] have investigated the way to generate Chinese traditional paintings by simulating the interactions between water, ink, paper and brushes. However, their performance is far from satisfactory compared to the state-of-the-art deep-learning models. In contrast, deep generative models have rendered reasonably good Chinese paintings. Wang et al [4] employ different generative adversarial networks to Chinese painting dataset created by themselves. Even though the

models show some promising results, the generated Chinese paintings suffer from mode collapse, where the outputs of the generative model have similar style despite a diverse set of styles in the training set. Also, the abstract aspect of the painting is lost due to lack of constraint. Li et al [1] propose a novel MXDoG-guided filter (**M**odified version of the e**X**tended **D**ifference-of-**G**aussians) to better preserve the abstract style of Chinese ink painting within the neural style transfer framework. The model offers some improvement over conventional style transfer methods in terms of aesthetic features (good textureless style, better contrast, etc). He et al [2] introduce three additional loss terms besides normal adversarial loss in GANs to constrain the style of voids, brush strokes, as well as ink wash tone and diffusion. The score of the stylization perceptual study improve after incorporating additional loss terms, and the generated ink wash paintings preserve impressive artistic style. Thus, additional terms in the loss function is the key to generate Chinese ink painting with desired abstract style. This is also one important part of our project.

3 Methods

In this section, we will introduce three different cGAN architectures and numerical evaluation metric we use in this project.

3.1 Conditional Generative Adversarial Networks

In the original generative adversarial networks, there's no control on modes of the generated data. However, in the conditional generative adversarial networks [8], by conditioning the model on additional information such as labels, text, images, we can direct the data generation process.

In our project, we will use the architecture of the Pix2Pix [15] model which is conditioned on input images as our baseline to generate Chinese ink paintings form sketches. In the first part, we will investigate the influence of various additional losses on painting generation and find the most suitable one for the Chinese ink painting style. In the second part, we propose to add category embedding into image embedding to generate Chinese paintings for different content categories. In the third part, we propose to add a latent random factor into image embedding aiming to generate diverse Chinese paintings.

3.1.1 Image-to-image translation (Pix2Pix)

As proposed in [15], Pix2Pix is a general-purpose codebase for pixel-to-pixel image translation, and is a variant of conditional GAN. This conditional GAN trains a generator G to learn a mapping from a subject image x and a random noise vector z, to target image y, $G : \{x, z\} \rightarrow y$. A discriminator D is trained as well, to distinguish the generated image from the real image. The training procedure is shown in Figure 2 [15].

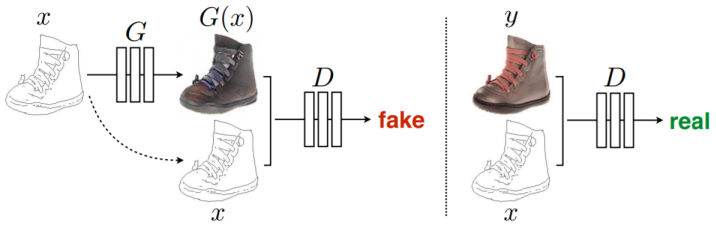

Figure 2: Training a conditional GAN to map edges $\rightarrow$ photo. The discriminator, D, learns to classify between fake (synthesized by the generator) and real {edge, photo} tuples. The generator, G, learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe the input edge map.

The objective function of this conditional GAN is as following:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[logD(x, y)] + \mathbb{E}_{x,z}[log(1 - D(x, G(x, z)))], \tag{1}$$

where G attempts to minimized the objective function, and D attempts to maximize it.

Pix2Pix adopts the generator architecture with skips in [16] which is called "U-net", to give the generator a means to circumvent the bottleneck of the encoder-decoder structure. Each skip simply concatenates all channels at layer i and layer $n - i$, where n is the total number of layers, which is shown in Figure 3 [15].

The discriminator adopts the PatchGAN architecture, which is very similar to convolutional nets, with details provided in [15].

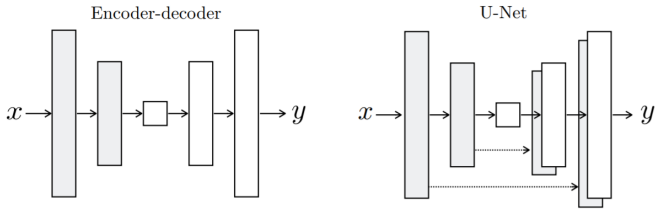

Figure 3: Two choices for the architecture of the generator. The "U-Net" is an encoder-decoder with skip connections bwtween mirrored layers in the encoder and decoder stacks.

As we mentioned before, directly applying Pix2Pix model may not be sufficient to generate stable and satisfactory Chinese ink painting style images. Therefore, we consider experimenting with the following losses to find the most suitable one for generating Chinese style paintings.

- **Pixel Loss** The simplest and most common loss is the average per-pixel difference between the generated image and the ground-truth image. Here we want to consider different pixel level losses: L1, L2, and SL1 (Huber). While L1 may be favored since it encourages sharp images and L2 encourages blurring images [15], it may not be the case with Chinese painting due to its unique ink style. It's also reasonable to consider the Huber loss, which can be seen as some combinations of L1 and L2 loss.
- **Style Loss** Inspired by the style transfer problem, we also consider the style loss defined as $L_s = \sum_{i,j}(G_{i,j} - A_{i,j})^2$ where G is the Gram matrix from feature map of the generated image, A is the Gram matrix from feature map of the target image and here we only consider the style loss of the last layer for simplicity.

3.1.2 Image-to-image translation with category embedding

One problem of the aforementioned Pix2Pix model is that it tends to map different categories into a single style when trained on mixed datasets such as mixed flower and mountain dataset. Inspired by [17], we propose to concatenate a non-trainable Gaussian noise as category embedding right before it goes through the decoder, by which the encoder still maps the same sketch image into the same image embedding vector, while the decoder, on the other hand, will take both the image embedding and category embedding to generate the target painting, as illustrated in the Figure 4. However, without additional constraints, the generator will still mix the categories together, generating paintings that don't look like any of the provided targets. Inspired by [18], we introduce the category loss to direct the discriminator to distinguish different categories by predicting the categery of the generated paintings. Therefore, with category embedding, we now have a conditional GAN that can handle multiple categories at the same time.

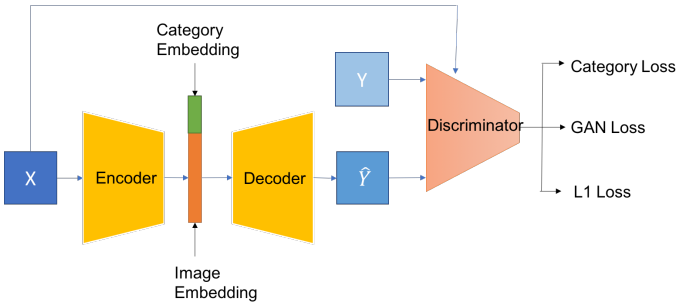

Figure 4: Framework of image-to-image translation with category embedding

3.1.3 Diverse image-to-image translation with category embedding

So far, the models we've considered are not capable of generating diverse paintings. To encourage diversity of the generated images, we propose to add a random latent code z which is from a standard Gaussian distribution to image embedding. Motivated by [5], we employ an additional encoder network that predicts the latent code from the generated image to capture the randomness, as illustrated in the Figure 5.

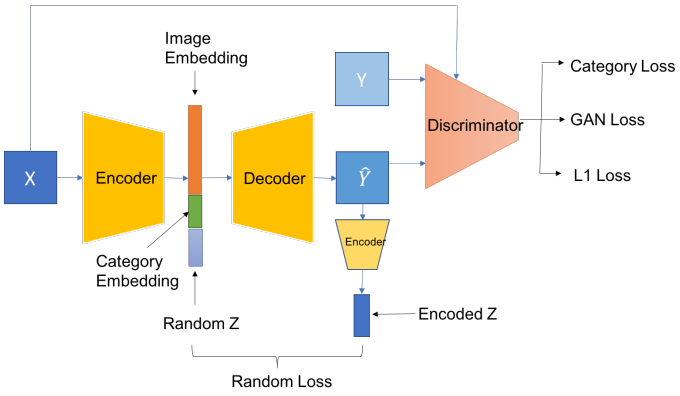

Figure 5: Framework of diverse image-to-image translation with category embedding

3.2 Evaluation Metrics

One challenge in this project is to quantitatively evaluate the quality of the generated images. For this issue, we consider assessing the similarity between the generated images with the real Chinese paintings. Some similarity metrics such as structural similarity (SSIM) [19] can be used in our project. SSIM measures the perceptual correctness by comparing contrast, luminance and structure of two images and it's invariant to scaling, rotation and insensitive to luminance and contrast change [20], which is suitable to evaluating outputs of our generated images from line sketches. SSIM is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_X^2 + \sigma_y^2 + c_2)},$$

where μ_x, μ_y are the averages of x and y, σ_x^2, σ_y^2 are the variances of x and y, σ_{xy} is the covariance of x and y, c_1, c_2 are two variables to stabilize the division with weak denominator.

4 Experimental Results and Discussion

We used the implementation of Jun-Yan Zhu's work on Pix2Pix model [21] as our baseline and developed our models based on it. To train our networks, we set up a virtual machine on the GCP using a Nvidia K80 GPU.

4.1 Preprocessing Data

We use existing dataset from Wang et al.'s work on Chinese Paintings Generation[4]. This dataset contains 5798 samples of Chinese painting each of which has 256x256 pixels. They are obtained by searching for specific keywords in Google and Baidu, cropping, and removing duplicates.

To fit into our needs, from this dataset we select 400 flower images and 400 mountain images respectively as training data and 30 images for each as testing data. We also search for a few real photos of flowers and mountains for test scenarios from Google. Then we apply Canny edge detection method [22] on the above images to generate their line sketches.

4.2 Qualitative and quantitative comparison of different losses on Pix2Pix

As we previously mentioned, directly applying cGAN to generate ink painting from sketch without any additional constraint may not give satisfactory results in that the abstractness of the painting is lost. Therefore,

we first investigate the influence of additional losses to the style of the generated paintings. We extend the Pix2Pix model to incorporate additional L1, L2, SL1, and style losses as mentioned in the methodology section.

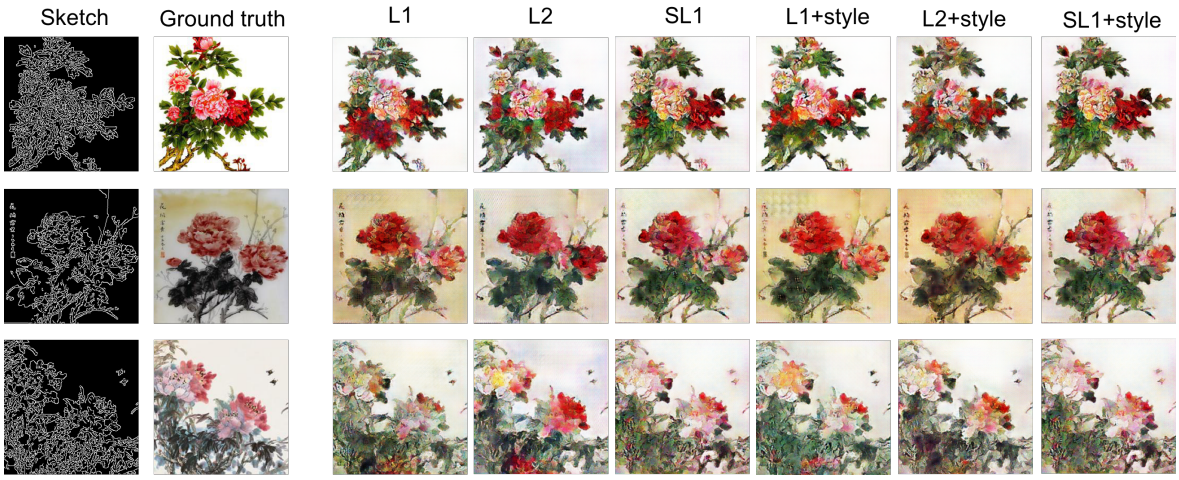

Figure 6: Qualitative comparison of different losses on image-to-image translation task

Figure 6 shows the qualitative comparison of different losses based on the baseline Pix2Pix model. Compared to the ground truth, paintings generated with L1 loss appears to be sharper and more realistic, whereas paintings with L2 loss tend to preserve some of the abstractness as well as the blurred features of the ground truth paintings. Also, when the style loss is added, the color of generated paintings matches quite well with the original one. This implies that L2 + style loss may give the best result. Indeed, this is supported by the quantitative comparison of different losses from SSIM, which can be seen from Figure 7. Therefore, applying additional L2 loss combined with style loss helps the generated painting preserve blurred features and color of the ground truth.

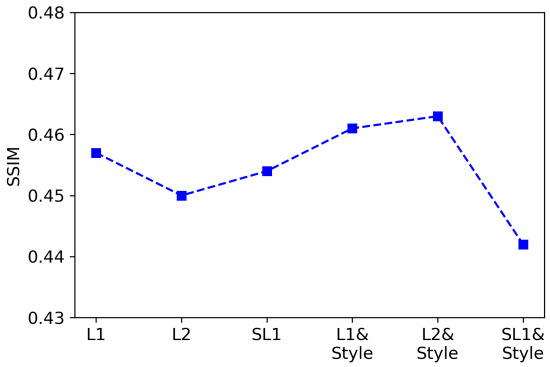

Figure 7: Structural similarity of different additional losses

4.3 Image-to-image translation with category embedding

So far, the baseline Pix2Pix model with additional losses can generate reasonably good ink paintings with Chinese style, but it only applies to the dataset with a single category, flower in the above case. It then follows that if we were to generate ink paintings of another category, we need to train a second cGAN on that category, which is extremely inefficient when the number of category becomes large. To overcome this issue, we propose a new cGAN architecture called Ske2Ink which builds on top of the Pix2Pix model, and incorporates category information into the encoded vector in the U-net. The generated ink paintings using Pix2Pix and Ske2Ink on a dataset mixing flower with mountain are shown in Figure 8.

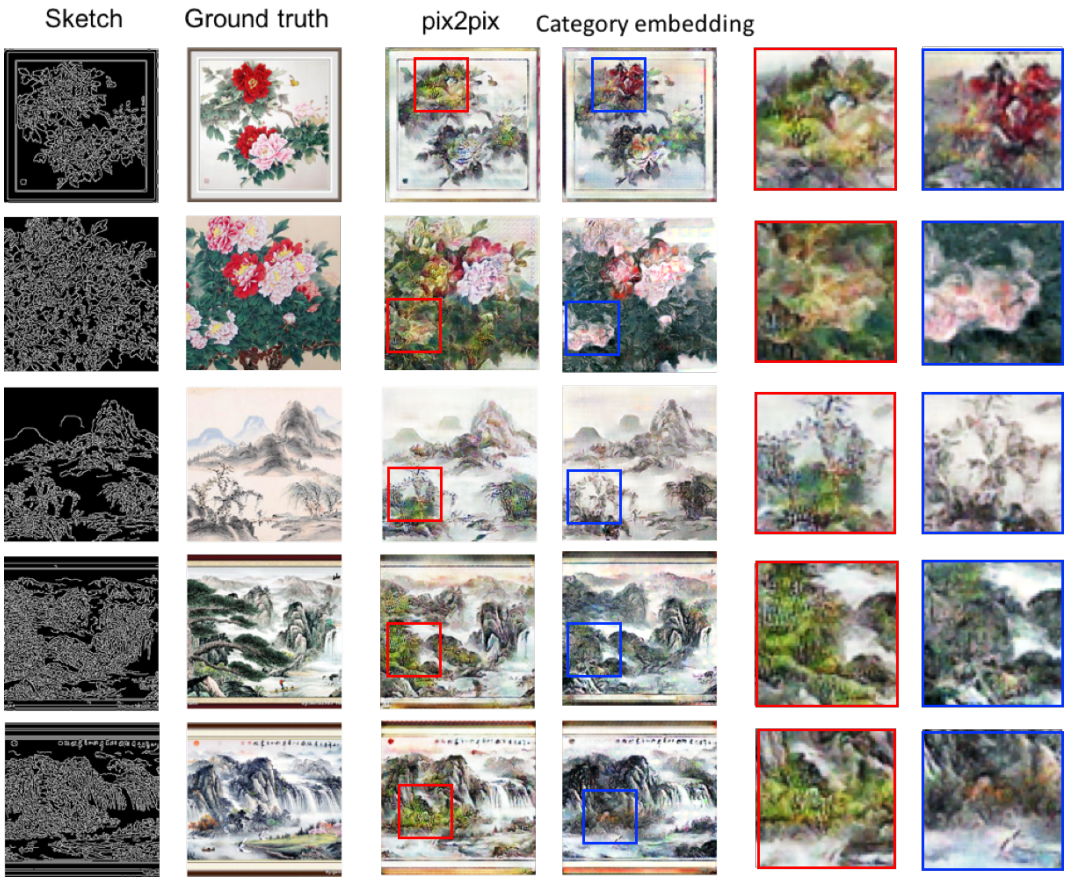

Figure 8: Comparison of Pix2Pix and our model with category embedding (Ske2Ink). The last two columns are the zoom-in of certain part of the generated images

When trained on the mix-category dataset, Pix2Pix model fails to distinguish the unique features of flowers vs mountains, and misinterpret flower as mountain, and vice versa, which is shown in the zoom-in part. In contrast, Ske2Ink model successfully captures these unique features, and generates ink paintings that match very well to the ground truth. Also, our model preserves some detailed brush strokes besides the color of the depicted objects as can be seen from the zoom-in, which is quite impressive. Thus, our Ske2Ink model outperforms the baseline Pix2Pix model on the mix-category dataset, and should be able to generalize to dataset with arbitrary categories in the subject matter.

4.4 Diverse image generation

With our proposed Ske2Ink model, we can achieve a good performance on Chinese ink painting generation task with a single style. It would be even more interesting if we can generate different style ink paintings from a single sketch. To do this, we further incorporate a latent random vector z into the image embedding in Ske2Ink, as illustrated in Figure 5. This random vector may provide some variations in the style of the generated paintings. However, we did not observe any randomness in style in the test scenario, which indicates that random vector z has little or no effect on the generator. In fact, when we look at the weights connected to z, they are extremely small after training. As a result, the generator completely ignores the randomness provided by z, which explains no randomness in the test time. We also employ a recent multi-modal image-to-image translation model called BicycleGAN [5] to see if a diverse style could be realized. From Figure 9, very little variation is observed in the generated paintings 1,2,4,5, whereas painting 3 completely fails to capture the correct style of the ground truth.

One possible solution to this mode-collapse problem in cGAN models would be explicitly regularizing the generator by additional maximization term in the objective proposed by Yang et al [23], which prevents the

Figure 9: Example generated paintings from BicycleGAN

norm of the gradient approaching 0 in the generator. This additional objective is easy to incorporate into the existing cGAN model, and should be a promising route for future exploration.

4.5 Painting generation from real photos

Finally, we further test our our model (Ske2Ink) with real flower and mountain photos to generate Chinese ink paintings. Some results are shown in Figure 10. In the test scenarios, we have several observations:

On the one hand, as shown in Figure 8, in the sketches of Chinese flower paintings, the flower itself usually contains dense edges, its leaves have less dense edges and the empty drawing paper has no edges. However, the sketches of real flower photos in Figure 10 do not quite match the above characteristics. Therefore in the generated paintings, we find some flower petals are painted as leaves or drawing paper, which makes the generated paintings not as good as real paintings. On the other hand, the mountain sketches from photos in Figure 10 basically match the characteristics of those from Figure 8. In the sketches, mountains always contain dense edges and places which have almost no edges are sky or clouds. This makes the difference between our generated paintings and real mountain paintings closer than those of flowers.

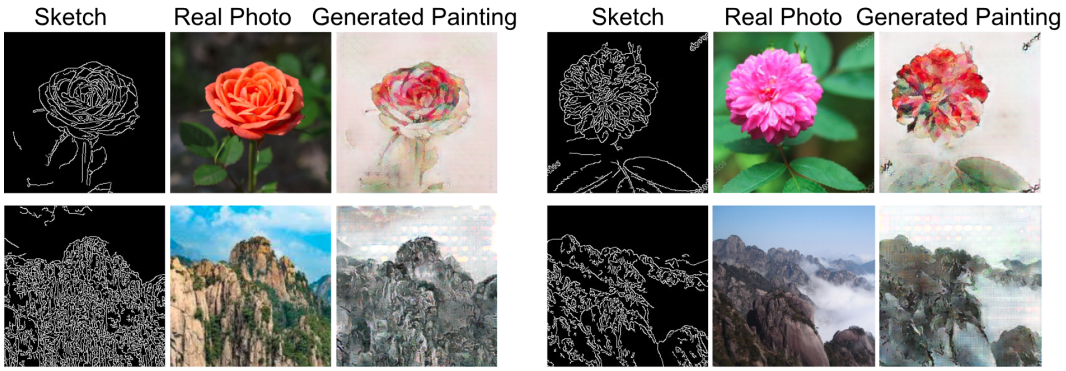

Figure 10: Paintings from real photos

Therefore, we could presumably believe that the Ske2Ink model learns the features from the sketches and colorize them accordingly, and there is a restriction on the style of sketches used to generated Chinese ink paintings, which must be followed when generating paintings of certain category.

5 Conclusion

In this project, we compare the influence of a series of additional losses on Pix2Pix model both qualitatively and quantitatively, and find the most suitable one for Chinese ancient ink painting style. We then propose and implement the Ske2Ink model to generate Chinese paintings for different content categories (flower and mountain), and it outperforms Pix2Pix which fails to capture the differences of these two categories. Moreover, our Ske2Ink model is capable of generating good quality ink paintings both from sketch figures and real photos. We also explore generating diverse Chinese paintings by adding randomness to image embedding, which does not perform well and requires further exploration.

6 Author Contribution Statement

C.M. and Z.D. designed the project. X.Z. and C.M. preprocessed the data. X.Z., C.M. and Z.D. developed the model architectures. Z.D. presented our work to audience. We all performed the experiments, analyzed the results and wrote the articles.

References

[1] Bo Li, Caiming Xiong, Tianfu Wu, Yu Zhou, Lun Zhang, and Rufeng Chu. Neural abstract style transfer for chinese traditional painting. *arXiv preprint arXiv:1812.03264*, 2018.

[2] 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 *ACM Multimedia*, 2018.

[3] 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*, 11:4, 2018.

[4] Yuan Chen Guanyang Wang, Ying Chen. Chinese painting generation using generative adversarial networks. 2017.

[5] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.

[6] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *CoRR*, abs/1601.06759, 2016.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[11] 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.

[12] SongHai Zhang, Tao Chen, YiFei Zhang, ShiMin Hu, and Ralph Martin. Video-based running water animation in chinese painting style. *Science in China Series F: Information Sciences*, 52(2):162–171, Feb 2009.

[13] Der lor Way, Yu ru Lin, and Zen chung Shih. The synthesis of trees chinese landscape painting using silhouette and texture strokes. *Journal of WSCG*, pages 499–506, 2002.

[14] Songhua Xu, Yingqing Xu, Sing Bing Kang, David H. Salesin, Yunhe Pan, and Heung-Yeung Shum. Animating chinese paintings through stroke-based decomposition. *ACM Trans. Graph.*, 25(2):239–267, April 2006.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

[18] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[19] Zhou Wang and Eero P Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–573. IEEE, 2005.

[20] Yağmur Güçlütürk, Umut Güçlü, Rob van Lier, and Marcel AJ van Gerven. Convolutional sketch inversion. In *European Conference on Computer Vision*, pages 810–824. Springer, 2016.

[21] Ingo Lütkebohle. CycleGAN and pix2pix in PyTorch. `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`.

[22] John Canny. A computational approach to edge detection. In *Readings in computer vision*, pages 184–203. Elsevier, 1987.

[23] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *CoRR*, abs/1901.09024, 2019.