Chinese Painting Style Transfer Using Deep Generative Models

Weijian Ma (krisma)
Microsoft AI and Research
krisma@stanford.edu

Yanyang Kong (yanyangk)
Microsoft AI and Research
yanyangk@stanford.edu

Abstract

Artistic style transfer aims to modify the style of the image while preserving its content. Style transfer using deep learning models has been widely studied since 2015, and most of the applications are focused on specific artists like Van Gogh, Monet, Cezanne. There are few researches and applications on traditional Chinese painting style transfer. In this paper, we will study and leverage different state-of-the-art deep generative models for Chinese painting style transfer and evaluate the performance both qualitatively and quantitatively. In addition, we propose our own algorithm that combines several style transfer models for our task. Specifically, we will transfer two main types of traditional Chinese painting style, known as "Gong-bi"(工笔) and "Shui-mo" (水墨)[9] to modern images like nature objects, portraits and landscapes.

1 Introduction

Neural Image style transfer is a fun technique for transferring a style from an image or one domain to another using neural network models. Style transfer using deep learning models has been widely studied since 2015. Most of applications afterwards are working on transferring the style from one specific style image to the target images, like transferring Van Gogh's *The Starry Night* to realistic photos. In 2017, Cycle-GAN came out as a novel approach to learn image translation from one domain to another without requirement for paired training data. In our project, we will experiment, evaluate Gatys's CNN based model and cycle-GAN based models on Chinese painting style transfer. Based on the experiment results, we propose our own approach which combines these two transfer models for Chinese painting style transfer.

Traditional Chinese painting known as "Guo-hua"(国画) is very different from Western styles of art. There are two main styles in Chinese painting, one is "Gong-bi"(工笔), which means "meticulous". It uses detailed brushstrokes to precisely depict figures and subjects. The other one is "Shui-mo"(水墨), meaning "water and ink", which is close to freehand style sketch using brush (check examples in Figure 1).

1.1 Related Work

In 2015, Gatys et al. [1; 2] first introduced an algorithm based on deep convolutional neural networks to create artistic imagery by separating and recombining content and style of images. In their work, they built a new feature space based on Gram Matrix statistics on the pre-trained layers of CNNs. It captures the image style by including feature correlations of multiple layers. Their key finding is that in CNNs, image content and style representations are separable. While they are able to show high perceptual quality, the most striking limitation is the algorithm is very slow since each inference requires solving an optimization problem. They also mentioned for image style transferring, evaluation criterion is neither mathematically precise nor universally agreed upon. In 2016, Johnston

CS236 (2019 Fall) Project Final Report
Code link: https://github.com/yanyangbaobeiIsEmma/ChinsePaintingStyleTransfer

(a) "Gong-bi"(工笔) Lotus

(b) "Shui-mo"(水墨) Mountains

Figure 1: Chinese painting examples, the left painting is in "Gong-bi"(工笔) style and the right side is a typical "Shui-mo"(水墨) painting.

et al. [3] proposed to combine per-pixel loss function and perceptual loss functions for training feed-forward networks for image transformation. They claimed the results of style transfer using their image transformation networks are qualitatively similar to Gatys et al.[1] but are much faster to generate. Generative Adversarial Nets for generating images was first proposed by Ian Goodfellow et al. [4] in 2014. The key idea of GANs is using competition between generative model and discriminative model to force generated images to be indistinguishable from genuine images. In 2017, Jun-Yan Zhu et al.[5] added cycle consistency loss on top of GANs by introducing an inverse mapping from target domain to source domain. Their algorithm made unpaired training possible for image domain translation. In their work for artistic collection style transfer, they trained model on landscape photos to take the style of the whole collection of artworks from Cezanne, Monet, Van Gogh, and Ukiyo-e respectively. Yuan Chen et al. [7] are the first to explore deep learning model applications on Chinese painting. In their project for CS 231N (2017), they tried out cGANs (conditional GANs), DCGANs (Deep Convolutional GANs) and WGANs (Wasserstein GANs) for Chinese painting generation. They took the edge of Chinese painting as input (with no color) and targeted to recover the colored version (the original painting). The way they trained GANs is still relying on paired training data, thus it cannot be directly applied to Chinese painting style transfer for realistic photos.

1.2 Overview of What We Do

In this project We start with preparing our dataset, which includes source dataset (realistic photos of people, flowers, birds, and landscapes), and target dataset (Chinese paintings including "Gong-bi"(工笔) and "Shui-mo"(水墨)). First we will examine the model introduced by Gatys et al.[1] in both "Gong-bi" and "Shui-mo" paintings. Then we train cycle-GAN with our own dataset in order to convert a realistic photo to Chinese painting. Since "Gong-bi" and "Shui-mo" differ from each another in their colors, narrative subjects, level of details and so on, we train these two categories separately. We will compare the performance of Gatys et al.[1]'s CNN model with cycle-GANs qualitatively and quantitatively, and discuss their pros and cons. Then we introduce our own three approaches of combing cycle-GANs and Gatys'CNN based model for Chinese painting style transfer.

2 Problem Statement

In our project, we want to transform a realistic photo like portrait, nature object, landscape to a Chinese painting in "Gong-bi" or "Shui-mo" style based on their topics using deep neural network models like examples in Figure 2.

(a) "Gong-bi"(工笔) Style Lotus

(b) "Shui-mo"(水墨) Style Mountain

Figure 2: Chinese painting style transfer examples, on left side we transfer a real lotus photo to a "Gong-bi" style Chinese painting using "Gong-bi" cycle-GAN, on the right side we transfer a mountain landscape photo to a "Shui-mo" style Chinese painting using naive combination model ("Shui-mo" CycleGAN + CNN neural transfer)

2.1 Dataset

Due to the fact that traditional Chinese paintings are scarcer (and harder to find) than realistic images, we first prepare traditional Chinese paintings (target dataset) from Chinese painting data base created by Yuan Chen et al.[8], and categorize them into "Gong-bi" and "Shui-mo" classes. In "Gong-bi" dataset, we have in total 512 paintings, which includes topics like human portraits, flowers, birds, etc. In "Shui-mo" dataset, there are 1796 paintings in total. Most of them are landscapes for mountains and forests. Then we prepare source dataset by scraping realistic images of the same narrative subjects through Google Image. we further resize the scraped images to 256 by 256 to match the Chinese painting dataset.

2.2 Evaluation Method

We will focus on three criteria for different style transfer models, 1) the ability to learn style representation perceptually, 2) inference speed (time complexity in inference) and 3) content loss, style loss defined in Gatys el al.'s paper. In this project, we will compare CNN neural transfer model, cycle-GAN and our own approach. We agree that style representation ability is hard to measure so we will compare the cycle-GAN output, CNN neural transfer output and output from our approach using the same source image side by side in section (4.1).

To measure the inference speed we will calculate the average time each model takes to generate one output image. Notice that Cycle-GAN is able to transfer output images in batch after the model is trained with whole dataset, while the CNN neural transfer model needs to optimize for each <source, target> image pair. We also compute and compare the content loss and style loss used in Gatys et al.'s paper [1] to measure the how close the generated image to the original content and the style.

3 Technical Approach

In this section, we will first briefly review Gatys's CNN neural transfer model and cycle-GAN models.Then we will describe our approaches to combine Gatys's neural transfer model with cycle-GANs.

3.1 CNN neural style transfer

We started with CNN neural style transfer model [1] as baseline model, which uses certain convolutional layers from pre-trained vgg-19 network trained on more than a million images from the ImageNet database. In Gatys et al.neural transfer model [1], it is required to have a pair of input images: style image $\vec{a}$ and content image $\vec{p}$, and the model will generate image $\vec{x}$ that mixes the content from the content image with the style form the style image. The loss function is defined by the sum of content loss and style loss

$$L_{total}(\vec{x}, \vec{a}, \vec{p}) = \alpha L_{content}(\vec{p}, \vec{x}) + \beta L_{style}(\vec{a}, \vec{x})$$

where α and β are the weighting factors for content and style reconstruction respectively. The content loss is defined as the squared-error loss between the feature representations of P^l from content $\vec{p}$ and

F^l from output $\vec{x}$ for each convolutional layer l,

$$L_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \left\| P^l - F^l \right\|^2$$

In the paper, they built a style representation that computes the correlations between the spatial features, which is given by Gram matrix $G(F) = FF^T$. For layer l, the Gram matrix G^l for feature matrix F^l is defined as $G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}$. And the style loss is the mean-squared distance between the Gram matrix from the style image's feature representation A^l and Gram matrix of the generated image's feature representation F^l at layer l, where N_l and M_l is the dimension of feature representation F^l.

$$L_{style}(\vec{a}, \vec{x}, l) = \frac{1}{4N_l^2 M_l^2} \left\| G(A^l) - G(F^l) \right\|^2$$

We review the performance of this model by randomly sampling pairs of content image (from source dataset) and style image (from target dataset) and apply style transfer. In general, the CNN neural transfer can learn the dominating color from the style image better than it learns other features like texture and shadow. However we also noticed a couple of limitations. First, we found that the model failed to separate style from content to some degree (check the example of CNN transfer model in Figure 3). Secondly, CNN neural transfer model needs to optimize parameters for each pair of content and style image, which makes the style transfer process, i.e., inference process very slow and hard to scalable.

(a) Source Image (b) Style Image (c) Output Image

Figure 3: Above is a CNN neural transfer example. In the output image, the woman's eyes and nose are in deeper color simply because the woman in style image is wearing darker outfit, and the outfit locates in the same region/coordinate of the image, which signals transferring of content instead of style.

3.2 Cycle-GAN

Cycle-GAN [5] is to learn a mapping $G : X \to Y$ such that the distribution of translated images $G(X)$ from source doamin is indistinguishable from the distribution Y, the targeted domain. It is also coupled with an inverse mapping $F : Y \to X$ and a cycle consistency loss to force $F(G(X)) \approx X$ and $G(F(Y)) \approx Y$. Since there are two mappings, there are two adversarial discriminators D_X and D_Y to distinguish between images from source domain to translated domain. In cycle-GAN, the loss function contains two types of terms: adversarial losses as GAN [4] and cycle consistency losses to prevent learned mappings G and F from contradicting each other. Thus the loss function in cycle-GAN is

$$L_{total}(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F)$$

where cycle consistency loss is defined as

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)} \left\| F(G(x)) - x \right\|_1 + \mathbb{E}_{x \sim p_{data}(y)} \left\| G(F(y)) - y \right\|_1$$

In our Chinese painting style transfer task, we trained two cycle-GAN models for "Gong-bi"(工笔) and "Shui-mo"(水墨) respectively.

3.3 Combined cycle-GAN with CNN style transfer (Our Approach)

After testing hundreds of samples using trained "Gong-bi" and "Shui-mo" cycle-GANs, we found that cycle-GANs are able to learn the general style representations very well while keeping the original contents of realistic photos. However, since they are trained on a big source domain, either "Gong-bi" general style or "Shui-mo" general style, when applied to a specific photo, the transferring performance for that specific topic/subject is not ideal (check examples in Figure 7). Here we provide three methods to improve cycle-GAN for Chinese painting transfer:

3.3.1 Naive combination

To combine the advantages of CNN transfer model (specific style learning) and cycle-GAN (quick inference), we could take the output from cycle-GAN as input image for CNN style transfer model, and choose original photo as content image and pick the style painting based on the subject in the photo (see Figure 4). The key difference is early stop in CNN transfer model, where we only take 10-20 iterations of parameter update of the 5 convolutional layers from pretrained vgg-19. Therefore the transfer process can be drastically sped up.

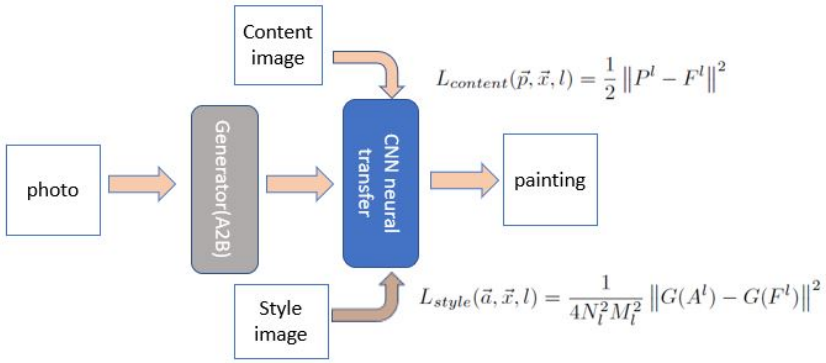


Figure 4: Naive Combination of cycle-GAN generator and CNN neural transfer model.

3.3.2 Use pre-trained cycle-GAN front CNN layers for CNN neural style transfer

Since the generator of cycle-GAN is a stack of convolutional layers, residual blocks and transpose convolutional layers (check the generator structure in Figure 5). We could use the pre-trained convolutional layers from generator for style and content representations like vgg-19 network in CNN neural style transfer model (section 3.1), where the pre-trained generator is from cycle-GAN models in section 3.2. We call this method "cycle-GAN neural transfer" in our paper. For this project, we tried this approach on both "Gong-bi" GAN and "Shui-mo" GAN.

3.3.3 Add style, content losses in cycle-GAN

We could also add style, content losses on the top convolutional layers of the generator from the pre-trained cycle-GAN, while fixing all other bottom layers (Check model structure in Figure 5). Given style image and content photo, we fine tune the top convolutional layers to reduce additional style and content losses. It is call "cycle-GAN fine tune" afterwards.

4 Experiments and Results

In CNN transfer model, we reused the setting as Gatys' paper [1] (using layer 'conv4_2' for content representation and layers 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1' and 'conv5_1' as style representations from pretrained vgg-19) and trained for 300 iterations for each optimization task.

In cycle-GAN models, We trained two cycle-GAN models for "Gong-bi"(工笔) and "Shui-mo"(水墨) respectively. In "Gong-bi" cycle-GAN model, our goal is to transfer realistic portrait or nature photos

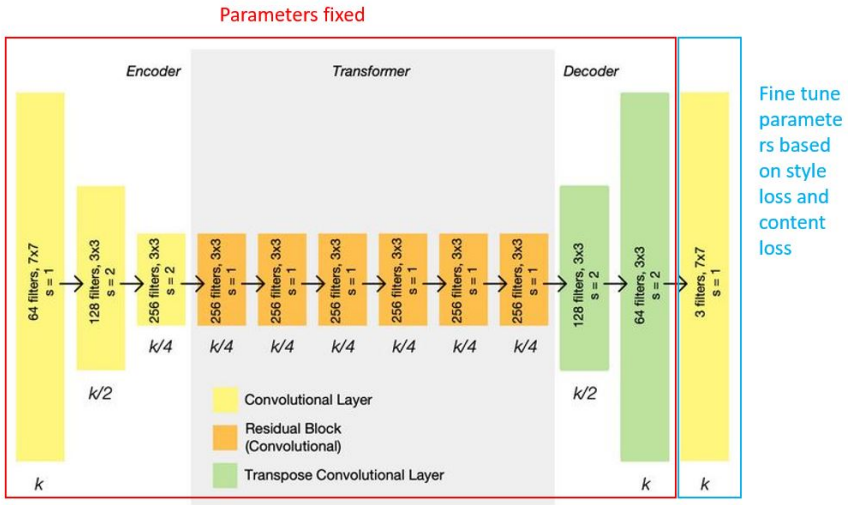
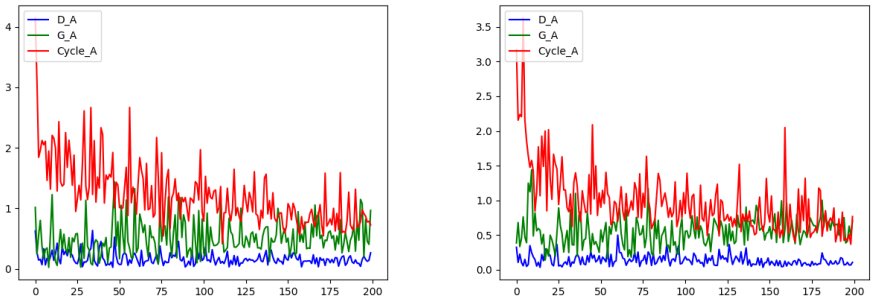


Figure 5: We add style, content losses for the generator in the top covolutional layers (in blue box), fix all parameters in the bottom (in red box), only fine tune parameters in the blue box.

to "Gong-bi" style. We prepared 4 categories photos in our source domain: human portraits, birds, lotuses, peonies, which are common subjects of "Gong-bi". In total, we have 477 photos with almost equal numbers in 4 categories for source domain, and 470 "Gong-bi" paintings for target domain. In "Shui-mo" cycle-GAN model, we used a larger dataset, which contains 1031 landscape photos for source domain and 1542 "Shui-mo" paintings in target domain.
We trained cycle-GAN models on single Nvidia GTX 2080 Ti GPU and we plot the losses from Generator $G : X \rightarrow Y$, discriminator D_X and the cycle consistency loss $E_{x \sim p_{data}(x)} \|F(G(x)) - x)\|_1$ during 200 epochs in Figure 6. It costs us $\sim$ 7 hours to train "Gong-bi" cycle-GAN model and $\sim$ 20 hours for "Shui-mo" cycle-GAN model but in inference time, it's just feed-forward pass for generator to generate stylish images, so it's a real time inference.



(a) "Gong-bi" Loss vs Epoch

(b) "Shui-mo" Loss vs Epoch

Figure 6: D_A stands for Discriminator loss from A to B and G_A stands for Generator loss from A to B, and Cycle_A is the cycle consistency loss from reconstructed A' to A.

In our own approach to combine cycle-GAN with CNN style transfer models, we take the trained "Gong-bi" cycle-GAN and "Shui-mo" cycle-GAN from previous method as inference step 1.
For 1) naive combination method, we take the generated images as input image for CNN neural model (vgg-19) with early stop (10 iterations compared to original CNN neural transfer using 300 iterations). In the meantime, we add original photo as content image, and pick the style painting based on the topic of the photo.

For 2) cycle-GAN neural transfer, we take the first three convolutional layers 'conv_1', 'conv_2', 'conv_3' for style representation and take the ReLu layer 'relu_3' after 'conv_3' for content representation from cycle-GAN generator. We take the output from cycle-GAN as input, use the original photo as content image, pick the style painting and train the neural transfer model using new style and content representations.
In 3) Cycle-GAN fine tune method, we add style, content losses on the final convolutional layer of the generator from the pre-trained cycle-GAN, fine tune its parameters to reduce the style loss and content loss while fixing other parameters in the bottom.

4.1 Model Perceptual Comparison

In Figure 7, we evaluated and compared the CNN neural transfer model, "Gong-bi" cycle-GAN and our three combination approaches on a realistic human portrait and a picked "Gong-bi" style painting . In Figure 8, we also evaluated and compared the CNN neural transfer, "Shui-mo" cycle-GAN and combination approaches on a mountain landscape photo with a picked Song (宋) dynasty "Shui-mo" style painting.

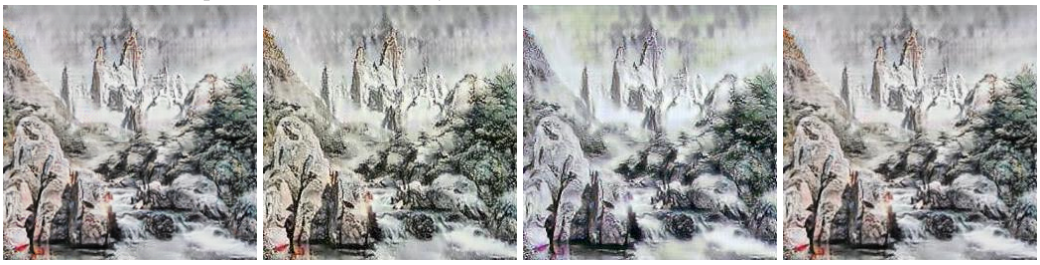
In both examples, we could see that naive combination gives better perceptual results, which also has near real time inference speed.

Figure 7: Model comparison on human portrait. CNN transfer output takes the background of style painting to output by mistake and also makes human face very vague. Cycle-GAN could generate sketch effect painting however with very weak "Gong-bi" style. Naive combination combines "Gong-bi" cycle-GAN and CNN transfer and provides clean human face with stronger "Gong-bi" style. Cycle-GAN neural transfer doesn't perform well compared to naive combination and cycle-GAN fine tune output is very close to the cycle-GAN's.

4.2 Inference Time Comparison

For inference time, CNN neural transfer model (section 3.1) takes around 20 minutes on CPU (about 200 iterations) for a <content image, style image> pair. For our three combination methods, "naive combination", "cycle-GAN neural transfer" and "cycle-GAN fine tune" take around 1-2 minutes on CPU (about 10 iterations) for any <content image, style image> pair. Since cycle-GAN only needs one forward pass for any input image, it's real time inference, which takes around 5 seconds.

Mountain landscape "Shui-mo" style CNN neural transfer



cycle-GAN naive combination cycle-GAN neural transfer cycle-GAN fine tune

Figure 8: Model comparison on landscape. CNN transfer did a good job by transferring the black ink wash style from "Shui-mo" painting to the mountain landscape but with fewer details on trees, water and rocks. cycle-GAN keeps the original colors of the landscape but also with very weak "Shui-mo" style. The naive combination could provide decent details of the landscape also transfer strong "Shui-mo" style to mountains. cycle-GAN neutral transfer gave additional ingredients to the output which looks nicer. And again cycle-GAN fine tune output is very similar to cycle-GAN direct output.

4.3 Style Loss and Content Loss Comparison

Here we compare the style loss and content loss for CNN neural transfer with our naive combination approach during the first 20 iterations for human portrait from Figure 7. In Figure 9, we could see that for naive combination, the style loss is much smaller than the style loss from CNN neural transfer model because we use the output from "Gong-bi" cycle-GAN as input image while the CNN neural transfer takes the original photo as input, which implies that cycle-GAN learns and captures the "Gong-bi" style though it is trained on a large "Gong-bi" source domain.

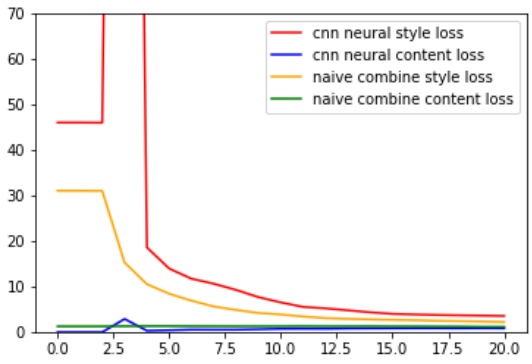


Figure 9: Style loss and content loss comparison for CNN neural transfer model and naive combination in the first 20 iterations. In application we only take 10 iterations for naive combination approach.

5 Conclusions

In this project, we implemented and evaluated Gaty's CNN neural transfer model, cycle-GANs for two Chinese painting style "Gong-bi" and "Shui-mo" transfer on realistic photos like human portraits, nature objects and landscapes. Based on that, we proposed three combination approaches to combine cycle-GAN and CNN neural transfer: "naive combination", "cycle-GAN neural transfer" and "cycle-GAN fine tune". We compared our methods with CNN neural transfer and cycle-GAN perceptually on some examples, and also compared their inference time and style loss, content loss during training. Based on our evaluations, "naive combination" provides high quality style transfer result and very short inference time (1-2 minutes). So we host this model (both "gong-bi" style and "shui-mo" style) on this website `http://20.185.103.117:5000/`. Try it out and have fun!

Acknowledgments

Thanks Yuan Chen, Guanyang Wang, Ying Chen for providing Chinese painting dataset.

References

[1] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge (2015) *A Neural Algorithm of Artistic Style*. `https://arxiv.org/pdf/1508.06576.pdf`

[2] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge (2016) *Image Style Transfer Using Convolutional Neural Networks*. `https://zpascal.net/cvpr2016/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf`

[3] Justin Johnson, Alexandre Alahi, Li Fei-Fei (2016) *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*. `https://arxiv.org/pdf/1603.08155.pdf`

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, YoshuaBengio (2014) *Generative Adversarial Nets*. `https://arxiv.org/pdf/1406.2661.pdf`

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (2017) *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. `https://junyanz.github.io/CycleGAN/`

[6] Xun Huang, Ming-Yu Liu, Serge Belongie1, Jan Kautz (2018) *Multimodal Unsupervised Image-to-Image Translation*. `https://arxiv.org/pdf/1804.04732.pdf`

[7] Guanyang Wang, Ying Chen, Yuan Chen (2017) *Chinese Painting Generation Using Generative Adversarial Networks*. `http://cs231n.stanford.edu/reports/2017/pdfs/311.pdf`

[8] Guanyang Wang, Ying Chen, Yuan Chen (2017) *Chinese Painting Dataset*. `https://github.com/ychen93/Chinese-Painting-Dataset`

[9] Wikipedia *Chinese Painting*. `https://en.wikipedia.org/wiki/Chinese_painting`