




# Interactively transforming chinese ink paintings into realistic images using a border enhance generative adversarial network

Chieh-Yu Chung<sup>1</sup> · Szu-Hao Huang<sup>2</sup> 

Received: 30 April 2021 / Revised: 18 March 2022 / Accepted: 15 August 2022 /

Published online: 27 August 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

Traditional Chinese painting has a long history. When we appreciate such paintings today, although we can obtain an overview of the landscape and environment of that time, it can be difficult to feel like we are interacting with the paintings. Alongside the rapid rise of deep learning, much research has been conducted on style transfer—for example, transforming photographs into the style of Chinese painting, sketches, or cartoons—but no research has considered the transformation of Chinese paintings into realistic images or even enriching such paintings through user interaction. To address this research gap, we employed a generative adversarial network (GAN), which is a generative model, to create new images that resemble the training data through the process of confrontation. Additionally, compared with general image-to-image translation, converting Chinese ink paintings into realistic images requires additional input because ink paintings contain texture and border features of relatively low quality. We combined cycle-consistent GAN with pix2pix and added a label function to establish a border enhance GAN with the purpose of enhancing the detail of border images and producing more accurate realistic images. In this manner, traditional Chinese paintings can be invigorated. Finally, we compared the image generated using our model with other benchmarks. The results revealed that the image generated using our model exhibited greater similarity to the actual photograph than did the benchmark images. Therefore, our model mitigates a major problem encountered in previous works and renders more realistic results. These interactive images clearly and profoundly convey Chinese culture, offering the user a novel art experience. Moreover, when viewers can interact with the input image by selecting different geologic styles, they can derive a relatively profound immersive experience. Our study can serve as a reference in transforming images (such as watercolor and oil paintings) with blurry borders.

**Keywords** Chinese painting · Deep learning · GAN · Interactive · Style transfer

---

✉ Szu-Hao Huang  
szuhaohuang@nycu.edu.tw

## 1 Introduction

Artificial intelligence (AI) has developed rapidly over the past few decades, and deep learning has been applied in many industries and fields, with various materials being analyzed and generated. Several new integrated AI and computer vision (CV) technologies—such as face changing [42, 47], facial recognition [16, 49], image classification [1, 6], and image style transfer—have been developed. However, the ability of computers to create art remains limited. Take paintings as an example; human painters can naturally or intuitively combine various elements, such as content and style, to create unique visual experiences. However, the results automatically generated by computers, notably in generative art studies that have employed evolutionary computation or other automated algorithms to produce visual art or compositions, still differ considerably from those of human beings. This is because abstract concepts such as style are difficult to understand by or to be defined for computers.

Thus, numerous studies on style transfer and with the goal of obtaining computer-generated art that resembles that of humans, such as in face style, landscape style, and comic-style transfer, have resulted in pictures with rich information being converted into unreal portraits with less information; well-known trained networks, such as VGG16, VGG19 [48], and the convolutional neural network (CNN) [27], have been employed in related studies. Researchers have tended to employ simulations of “color” and “texture” to accurately achieve a certain style of painting; thus, the content and style of a picture are typically extracted from Content Image and Style Image, respectively. However, this method is limited to the conversion of real images into a painting; in this process, the style transfer effect is achieved by discarding some information or adding some styles to images with rich information. Therefore, not all materials and transformations are suited to this method. For example, converting Chinese ink paintings into realistic images is challenging because such paintings contain relatively low-quality information. There is a fundamental difference exists between Chinese ink paintings and other images. First, due to the skills and tools employed in ink paintings, the borders between objects are not clear in such paintings. Transforming images with blurry boundaries results in the generation of a vague texture output, which differs from the real world in which objects have distinct boundaries. Second, compared with other paintings (such as single object sketches), ink paintings include a more complex composition. Considering these two points, we must outline the borders between objects in Chinese ink paintings. In recent years, several studies have focused on the conversion of paintings into real images by using self-supervised learning algorithms. Liu et al. [36] proposed an unsupervised method that incorporates an attention module and self-supervised denoising to deal with abstractions and style variations that are specific to sketch images. B. et al. [35] presented a self-supervised AutoEncoder that can decouple style and content features to efficiently synthesize sketches for RGB-only datasets. However, the process of transforming Chinese ink paintings into realistic images not only requires additional information but also necessitates the consideration of the limited information in hand paintings; therefore, comparing such paintings with real images is challenging. In addition, we show the result using sketchy-to-image methods based on Liu et al. [36] and compare it with the original Chinese ink painting as Fig. 20. As the result shows, these kinds of methods cannot make a good transformation.

In addition, Most of the research on Chinese paintings has differed our research; scholars have transformed realistic images into Chinese ink paintings, but we transformed Chinese ink paintings into realistic images. Moreover, scholars investigating general graphics

generation have not performed similar research. Xue [56] employed generative adversarial networks to generate Chinese landscape paintings. Chen [4] designed and improved a Chinese ink painting rendering algorithm by using a deep learning framework and the convolutional neural network model. Lin et al. [34] proposed a multiscale deep neural network for transforming sketches into Chinese paintings. They trained their generative network by using both L1 loss and adversarial loss to generate more realistic images. However, few other researchers have transformed realistic images into Chinese paintings.

Chinese ink painting is valuable, but few people have noticed the potential of using Chinese ink as the main material in art. Ancient Chinese artists used brushes, soft pens, and their fingers as tools to create wonderful ink paintings. They recorded beautiful landscapes, humanity, the arts, and quotidian life to express themselves. Moreover, the traditional method of painting with ink on silk or rice paper is an ancient eastern academic category. Through these paintings, precious cultural heritage can be understood, the human landscape of the time can be experienced, and self-cultivation and enrichment of cultural understanding can be achieved. In particular, direct interaction with Chinese ink paintings can breathe new life into these paintings and enable users to express themselves through this process, enabling them to gain experiences that could be acquired by traveling back in time to thousands of years ago. Therefore, one of the major issues we wanted to address in this research was providing an interactable machine learning approach, which is targeting to generate a realistic style image. Since one of the core values of art is the creativity of humans, users should be more actively involved in the process of the creating process. Zhou and Wang et al. [58] proposed a sketch-to-Chinese painting approach that allows users to input their personal creations and generates satisfied “Shanshui” painting documents. Cheng and Gan et al. [5] introduce an agent that takes natural language description, targeting to edit images through sequential and textual commands. Currently, an increasing number of interactive and generative applications aim to increase art–people interactions to enhance the exhibition experience.

Therefore, this study aimed to transform Chinese ink paintings into realistic pictures that enable user interaction by employing a neural network. Our research goal was not only to enhance the understanding of research related to Chinese ink painting but also to provide people with a novel art experience by enabling direct interaction with Chinese ink paintings. Art lovers can express themselves through this process and imagine traveling back in time thousands of years and standing before the astonishing mountains.

To achieve our goal, we created a border-enhancing generative adversarial network (BEGAN) architecture for image-to-image translation of Chinese ink paintings. Our framework can be divided into two parts. The first part compensates for the considerable difference in edge structures between hand-painted and realistic images, we used a cycle-consistent generative adversarial network (CycleGAN) to first determine the details of real photographs—such as the texture of mountains, the shadows in landscape photographs, and colors—and added the details to the border map of Chinese ink paintings; in this manner, the photographs resulting from the transfer were more realistic. To the best of our knowledge, we are the first to focus on enhancing the borders and landscape details of Chinese ink paintings for image-to-image translation. In our second part, we input the results of the supplemental information into pix2pix to generate a realistic image in addition to adding labels to the image together with the user’s mark. Because mountains and skies are the two largest elements in our input images, we decided to ignore small objects and labeled the images by separating only mountains and skies. The object labels are concatenated with the generated border image; this layout can be utilized to enhance user engagement in image

generation. Border labeling is crucial for performing object identification in our model. If we use only border images with no labels as the input, the generator may fail to determine correctly whether the block is to be generated into a sky or a mountain. The purpose of this method was to provide the generator with the boundary of the input image as well as the label information to facilitate the division of the learning scope and target. This step also served as the initial interaction between the user and the architecture.

Moreover, to maximize the customization of the image results, we established a labelling approach for the user. On the premise of not destroying the harmony of images, the user can add some of their own ideas and creativity to the images. For example, the user can add snow to mountains by labelling the generated image, making their results different from others'. User engagement through direct input can enhance the generated result, thereby satisfying the user's imagination and demand for involvement.

The major contributions of this study are as follows:

- The conversion of original Chinese ink paintings with limited details into realistic images is achieved by enhancing the details of the border map using CycleGAN.
- Labels are concatenated to the enhanced border map when passed into pix2pix; in this manner, the generator can divide the learning scope and target more clearly.
- Providing a reference in transforming images (such as Chinese paintings) with blurry borders.
- The user can identify the area on the photograph that they wish to adjust; this interaction enables the user to be more than a passive recipient: they collaborate with the system on the new image, which makes every creation unique.

This paper is organized as follows: in Section 2, we review related works, and in Section 3, we introduce our methods and architectures. In Section 4, we present the experimental results and explain how we arranged our training set. Finally, in Section 5, we conclude the paper, discuss our research contributions, and provide future research directions.

## 2 Literature review

This section reviews two bodies of literature: (1) that on unpaired image-to-image translation performed using CycleGAN networks and (2) that on style transfer.

### 2.1 Unpaired image-to-image translation using CycleGAN

A generative adversarial network (GAN) [14] is a framework for estimating a generated model through the process of confrontation. It requires the training of two models simultaneously: one is called the generator (G), which is for learning the data distribution; the other model, the discriminator, is used to determine whether the sample is generated by the generator or taken from the training data. The outputs of the generator must be as real as possible to be able to fool the discriminator. The inputs for the discriminator have two sources: one is the outputs of the generator, and the other is real images. The job of the discriminator is to recognize the real images from all data inputs; that is, it must distinguish between real images and those generated by the generator. In this manner, the two models are continually optimized through the adjustment of their parameters, and eventually, the generator produces life-like images with high quality.

CycleGAN [60] can be regarded as a diversification of the conventional GAN. The major difference between them is that CycleGAN has more than one generator and discriminator, which enables the architecture to transfer styles between two input domains. It learns to capture the most representative characteristics of the source image and attempts to determine how those characteristics can be translated into the target domain by combining the features of the source image and those of the target image; this results in an image that simultaneously has elements of the transfer of domain A into domain B and that of domain B into domain A.

The transfer of an image from domain A into domain B—for example, transferring colors to a grayscale image, semantic labels to real photographs, and border images to pictures—is referred to as image-to-image translation [21]. Numerous studies on CV, image processing, and graphics have employed pair images to create a translation system [17, 21, 24, 54]. Nevertheless, pairs of data that can serve as training data are in short supply. Thus, learning from unpaired data has become a crucial topic. CycleGAN is an algorithm that aims to determine the translation of images between different domains use of without a paired dataset. It assumes a relationship between the two domains and attempts to create the connections between them. The system operates as follows: first, it is given a set of images that is in domain  $X$ , and a different set of images in domain  $Y$ . The training goal is mapping  $G : X \rightarrow Y$ , to result in the output  $\hat{y} = G(x)$ , where  $x$  is the image in domain  $X$  that is so similar to the images in domain  $Y$  that the discriminator cannot distinguish between  $y$  and  $\hat{y}$ . That is, following this training process, the distribution of  $y$  matches the empirical distribution  $p_{data}(y)$ .

However, using a unidirectional procedure could not guarantee that the output  $\hat{y}$  and input  $x$  become a meaningful pair; that is,  $G$  has infinite means to learn a mapping that has the same distribution between  $y$  and  $\hat{y}$ . To avoid this type of error, the structure exploits the fact that the translating system must be cycle consistent; that is, an image is translated from a horse to a zebra, and then it must be translated back from the zebra to the original horse. This means two generators are employed,  $G : X \rightarrow Y$  and  $F : Y \rightarrow X$ , where  $G$  and  $F$  should be the inverse of each other, and the mappings should be bijections. By training  $G$  and  $F$  simultaneously and adding the cycle-consistency loss [59], the goal of obtaining  $F(G(x)) \approx x$  and  $G(F(y)) \approx y$  can be achieved, and the adversarial loss and cycle-consistent loss can be summed to obtain the objective of unpaired image-to-image translation.

In recent years, several limitations of the original CycleGAN have been identified, such as the poor quality of generated images, unsuitability for asymmetric mapping, and susceptibility to noise. Thus, improvements to CycleGAN have been proposed, including Embedded CycleGAN [37], with its capacity to learn additional information for adaptive shape changes in unpaired image-to-image translation. Chen et al. [3] extended CycleGAN to include a quality-aware GAN (QGAN) framework to enhance the quality of generated images. Asymmetric CycleGAN architectures [8, 31] have been proposed for asymmetric translation, and Jia et al. [22] recommended a Lipschitz regularized CycleGAN to improve semantic robustness. Dai et al. [33] introduced a selective transfer cycle GAN to boost the training efficiency in unlabeled target domain. Dai et al. [7] improved the colorization quality of images by applying the perceptual loss and the total variation (TV) loss.

Relevant papers have revealed that GANs have diverse uses, but they are mostly used for generating images. However, their architectures are not suitable for every material, especially Chinese paintings, which have rarely been studied in this context. As mentioned previously, a conventional GAN can not only generate images from noise but also achieve

style transfer. Considering the limitation of pix2pix translation, which is also a style transfer method using a GAN model, Lu et al. [38] presented a modified generative model that can learn the joint distribution of a sketch and the corresponding image through the employment of joint images; this can thus ameliorate the effect of strict alignment requirements stipulated in the translation process. Osahor et al. [39] adopted a GAN model that uses a hybrid discriminator that can effectively classify multiple target attributes to improve the generation quality, thus synthesizing several synthetic images with special attributes from a single sketch image. In recent years, most studies [50] have addressed the transformation of images from real photographs into other styles, such as comics and sketches. However, few studies have been conducted on the conversion of paintings into realistic images, which is challenging due to the requirement of additional information and the difficulty associated with comparing paintings with real images. In our case, this means transforming the image from a Chinese painting style into the style of a realistic image.

## 2.2 Style transfer

Numerous papers have claimed to propose a method for transferring the style of images. In art circles, people define the style of art, especially paintings, by determining its color and texture; these two elements are thus the main factors considered when transferring the style of an image. The Visual Geometry Group at the University of Oxford created a now prominent network called VGG [48]. This network was initially used for image recognition; however, more recently, numerous studies have employed it for style transfer or even extraction of the first  $N$  layers to achieve their goal. This model is popular because it was trained on large amounts of data and uses multiple convolution layers in processing; the pretrained model has become a generic, flexible solution. In addition to VGG's application in image recognition, a well-known paper published by Gatys et al. further used VGG to achieve style transfer with their neural algorithm of artistic style [12]. On the basis of this concept, we posited that a CNN can be employed to capture features and thus achieve style transfer. Researchers have applied diverse approaches, for combining an art image and a real photograph to transfer the style of the photograph [12, 23, 26, 57], to transfer a facial photograph into a sketch photograph [2, 41] or a real photograph into a comic-style image [43, 55], or even to transfer artistic style in real time, such as the style of videos [11, 24, 45, 53]. Khan et al. [25] proposed an CNN with multi-convolutional-learning technique for photographic painting style transfer. To accelerate the process of high-resolution image style transfer, the super-resolution style transfer network (SRSTN) [32] has been proposed to stimulate the operation and reduce the memory usage at run-time.

Most related papers tend to transfer images from real photographs into styles, such as comics and sketches. In our special case, we were faced with the great disparity in information between realistic images and Chinese paintings. To address this problem, we had to modify the information in the two types of image to be more similar, enabling the architecture to learn more features; using this approach, the results were less blurry and more closely resembled real photographs.

## 3 Proposed method

In this section, we first provide an overview of our system. Subsequently, we separate our system into two parts, introduce each in detail, and implement the procedures step by step. First, we describe how we improved and modified CycleGAN [60] to become a suitable

part for our system. Then, we introduce the pix2pix [21] step by step, including the implementation method and the details of the model. Most importantly, we expanded pix2pix to fit our research goal and focus on the object to be generated. The resulting cyc-pix system is used for transforming Chinese paintings into real photographs.

### 3.1 System overview

In this section, we first briefly provide an overview of our system; subsequently, we introduce our system in detail. Figure 1 displays an architecture graph of our system, demonstrating the process from the original painting to final user-generated image. First, we converted our dataset into a border map by using the Sobel method to enhance the image detail. In addition to border images, we converted 51 images into label maps to guide the network regarding the positions of the skies and mountains. In the final step of data preprocessing, we selected the representative and clean images as the training set. Subsequently, we performed four experiments—enhancement, color rendition, BEGAN, and user interaction—to transform the input data into realistic and interactive images.

As Fig. 1 reveals, the process is separated into four steps:

**Chinese ink painting → border image** We employ the Sobel method, commonly used for edge detection, to obtain the first-order gradient of a Chinese ink painting and transform it into an image with borders.

**Border image → enhanced border image** In the first part of the border enhance generative adversarial network (BEGAN), we enhance the details of the border images to generate a clearer input for the second part of the BEGAN.

**Enhanced border image → realistic image** In the second part of the BEGAN, the user has their first interaction with the image; they can select their favorite style and create an image according to that style.

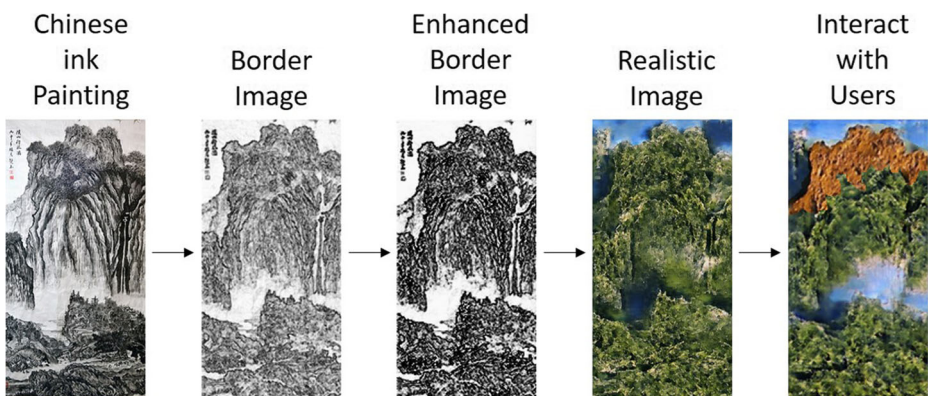


Fig. 1 System overview

**Realistic image → user interaction** Once the first version of the image has been generated, the user can draw on specific areas of the image with their preferred style (e.g., drawing some snow onto the top of Ali Mountain); in this manner, the final result is customized and meets the user's expectations.

### 3.2 Border enhancement method

Our research focuses on transforming Chinese ink paintings into realistic photographs by not only retaining the original cultural features but also adding a novel style. The goal is to provide an audience with a fresh means of enjoying Chinese ink paintings. After analyzing many such paintings and comparing them with real photographs, we noticed that some details cannot be depicted in Chinese ink painting.

Thus, to compensate for the considerable difference in the distinctiveness of edge structures between hand-painted and realistic images, we first created a structure that is used for enhancing the border image of Chinese ink paintings, with special emphasis on the aspects where Chinese ink painting differ the most from photographs—for example, clarifying details of mountains, enhancing the texture of the image to maximize the visibility of the border and facilitate separation of the border into different objects, and even erasing extra textures that may confuse the network (these textures are produced by the material of Chinese ink paintings and will not appear in realistic images). If the label information is not used as the input in our model, the generator cannot ascertain whether the block belongs to the sky or mountain. In other words, because of the typically blurry edges in ink paintings, the generator may not be able to distinguish between blue skies and green mountains. To address this problem, we add labels to the border image to provide information to the generator about the position of the sky and the mountain in the painting. Another goal of the border enhancement model was to erase lines that would negatively affect the results in order to make the border image resemble the real photo to a greater extent. Because we set only two labels in the data, our network considered every object to be a mountain.

Because the model architecture incorporates a border enhancement technique, BEGAN achieved the best performance in transforming Chinese ink paintings. In contrast to sketchyGAN, our model is specialized for transforming pictures with blurry edges and may not perform well at transforming images with other styles.

#### 3.2.1 CycleGan

This type of design introduces the concept of cycle consistency [60]; the architecture functions like a cycle. Unlike the conventional GAN, which contains only  $G : X \rightarrow Y$ , this architecture adds another generator,  $F : Y \rightarrow X$ , to the original architecture to ensure that the generator can transfer the content not only from domain  $X$  to domain  $Y$  but can also transfer the generated result back to domain  $X$ . Because we employ two generators, we also designed two discriminators,  $D_X$  and  $D_Y$ .  $D_X$  is used to determine whether the image is from the source domain  $X$  or is a translated image  $F(y)$ , which is generated from  $F$ .  $D_Y$  uses the same method to distinguish between  $Y$  and  $G(x)$ .

The training goal was to obtain a converted image that was as close as possible to the original image, thus enabling a corresponding image to be generated in the target domain. Therefore, when implementing training, we added another symmetric structure to enable two cycle-consistency losses.



### 3.2.2 Loss function

Because we could not obtain a predicted value that exactly matched the real value, we designed a function to estimate the discrepancy between those values, called the loss function. The loss function is used to determine the differences between the true value and predicted value. The ultimate goal is to obtain the smallest possible loss; the smaller the loss, the more robust the model.

**Adversarial loss** This architecture uses the adversarial loss [14] in the two mapping functions, which is the same as the loss in the conventional GAN. In the mapping function  $G : X \rightarrow Y$  and its discriminator  $D_Y$  we can define the objective function as follows:

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[\log D(y)] + \mathbb{E}_{x \sim p_x(x)}[\log(1 - D_Y(G(x)))]. \tag{1}$$

$G$  attempts to generate images  $G(X)$  that are as similar as possible to the images in domain  $Y$ , and the purpose of its discriminator  $D_Y$  is to determine whether the images are from the translated sample  $G(x)$  or from the real sample  $y$ .  $G$  aims to minimize the targeting of  $D$ , which is trying to maximize it; that is,  $\min_G \max_{D_Y} \mathcal{L}_{GAN}(G, D_Y, X, Y)$  [60].

This architecture also uses a propinquity adversarial loss for the other generator  $F : Y \rightarrow X$  and its discriminator  $D_X$  in the same manner ; that is,  $\min_F \max_{D_X} \mathcal{L}_{GAN}(F, D_X, Y, X)$ .

**Cycle-consistency loss** In theory, an adversarial training network can learn to map  $G$  and  $F$ , which respectively produce the same distribution of output as the target domains  $Y$  and  $X$  (strictly speaking, this requires that  $G$  and  $F$  are stochastic functions) [13]. However, with sufficient capacity, the network can map a set of input images of the same capacity to any random permutation of images in the target domain, where any learning map can obtain an output distribution that matches the target distribution. Therefore, one adversarial loss cannot ensure that the function that was learned can map a single input  $x_i$  to the expected output  $y_i$ .

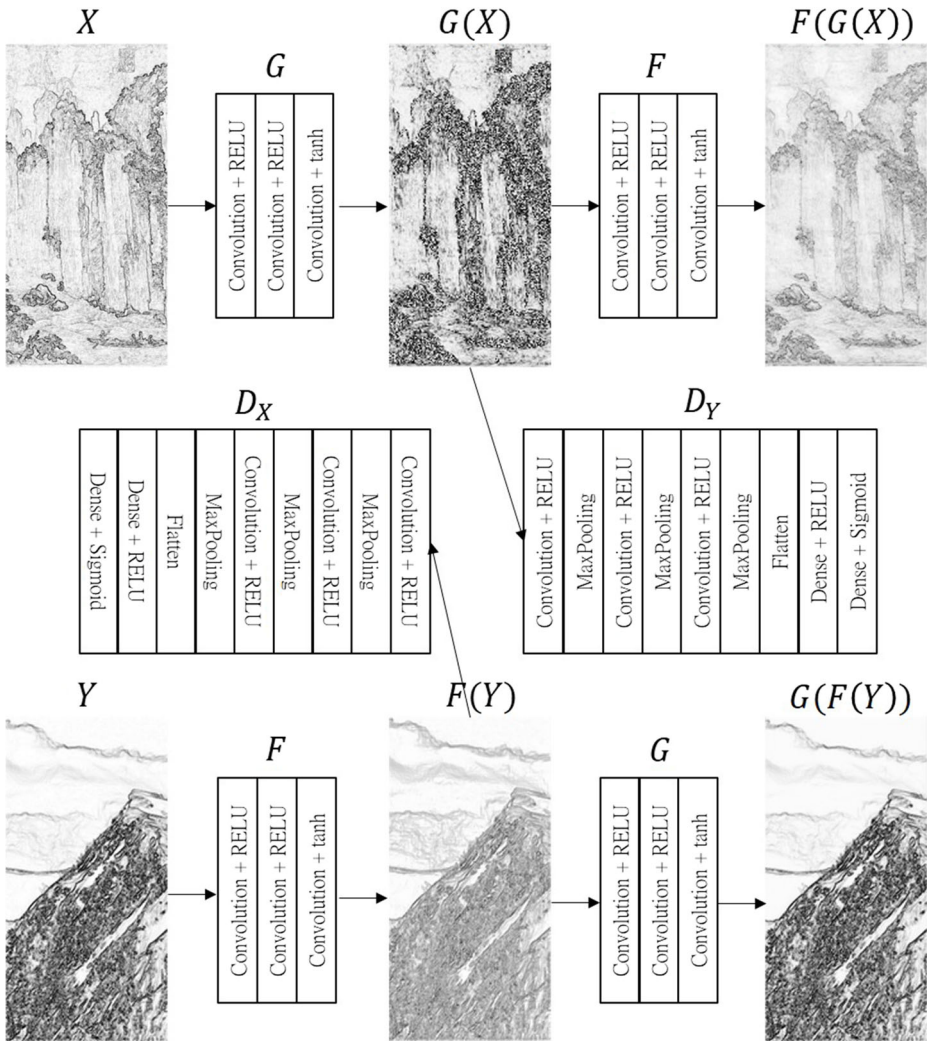
Figure 2 depicts the architecture of the CycleGAN. The system attempts to translate every image  $x$  in domain  $X$  into domain  $Y$ , and the translation cycle must also transform the translated images back into the original image; that is,  $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$ . This mechanism is called *forward cycle consistency*. Similarly, every image  $y$  in domain  $Y$  should meet this criterion—that is,  $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ , which is called *backward cycle consistency*. In sum, this architecture has *cycle consistency loss* [60] as follows:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1]. \tag{2}$$

**Full objective function** To mix the aforementioned objective functions, we sum them to obtain the full objective function:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F), \tag{3}$$

where  $\lambda$  is used for controlling the relative importance of the each objective function. This architecture was used to train the two types of cycle consistency simultaneously by adjusting



**Fig. 2** Architecture of the Cycle-GAN. The model contains two mapping functions  $G : X \rightarrow Y$  and  $F : Y \rightarrow X$ , and their discriminators  $D_Y$  and  $D_X$

the parameters of the generators and discriminators. In this manner, we address the min–max problem:

$$G^*, F^* = \arg \min_{G, F} \max_{D_x, D_y} \mathcal{L}(G, F, D_x, D_y). \tag{4}$$

Algorithm 1 optimizes (4) to obtain the desired result.

### 3.3 Enhanced border to realistic image with object labelling

Instead of the current model, pix2pix [21], we created a BEGAN, which has a structure based on pix2pix and is more appropriate for our question. The most crucial requirement in our structure is to not only generate images with correct shapes but also correct objects.

- 1: **for** number of training epochs **do**
- 2:     **for**  $k$  steps **do**
- 3:         Sample minibatch of  $M\{x^{(1)}, x^{(2)}, \dots, x^{(M)}\}$  from image set  $X$  without replacement.
- 4:         Sample minibatch of  $M\{y^{(1)}, y^{(2)}, \dots, y^{(M)}\}$  from image set  $Y$  without replacement.
- 5:         Synthesize  $M$  synthesized  $Y$  images  $\{G(x^{(1)}), G(x^{(2)}), \dots, G(x^{(M)})\}$  using the incumbent  $G$ .
- 6:         Synthesize  $M$  synthesized  $X$  images  $\{F(y^{(1)}), F(y^{(2)}), \dots, F(y^{(M)})\}$  using the incumbent  $F$ .
- 7:         Update the discriminator  $D_Y$  by ascending its stochastic gradient:
- 8:

$$\nabla_{\theta_{D_Y}} \frac{1}{M} \sum_{i=1}^M [\log D_Y(y^{(i)}) + \log(1 - D_Y(G(x^{(i)})))]$$

where  $D_Y(y^{(i)}) = True$  and  $D_Y(G(x^{(i)})) = False$   
for  $i = 1, 2, \dots, M$ .

- 9:         Update the discriminator  $D_X$  by ascending its stochastic gradient:
- 10:

$$\nabla_{\theta_{D_X}} \frac{1}{M} \sum_{i=1}^M [\log D_X(x^{(i)}) + \log(1 - D_X(F(y^{(i)})))]$$

where  $D_X(x^{(i)}) = True$  and  $D_X(F(y^{(i)})) = False$   
for  $i = 1, 2, \dots, M$ .

- 11:     **end for**
- 12:     **for**  $h$  steps **do**
- 13:         Update the generator  $G$  and  $F$  by descending their stochastic gradient:
- 14:

$$\nabla_{\theta_{G, \theta_F}} \frac{1}{M} \sum_{i=1}^M [\|F(G(x^{(i)})) - x^{(i)}\|_1 + \|G(F(y^{(i)})) - y^{(i)}\|_1]$$

where  $x^{(i)} \approx F(G(x^{(i)}))$  and  $y^{(i)} \approx G(F(y^{(i)}))$   
for  $i = 1, 2, \dots, M$ .

- 15:     **end for**
- 16: **end for**
- 17: Where  $\theta$  represents network parameters, and the gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

---

**Algorithm 1** Cycle-consistency generative adversarial nets using minibatch stochastic gradient descent for training. The number of steps to apply to the discriminator,  $k$ , and apply to the generator,  $h$ , are hyperparameters. We used  $k = 1$  and  $h = 3$  in our experiments, and it can be adjusted to fit different questions.

Take our material as an example; the objects in our image are the sky and a mountain, which means that the image contains more than one object. If we only use the border image without any labels as the input, the generator will be unable to determine whether the edge should be converted into sky or mountain. That is, with this type of blurry information, the generator may generate a green sky and a blue mountain. To avoid such mistakes, we attempted to add labels to the image, informing the generator of the location of the sky and the mountain. With these labels, the generator knows that the image contains different objects, and through the labels, it can generate objects according to the edge and label. Moreover, we utilize these labels to establish a user-driven labeling approach. The user can modify the label map and generate the desired images under the condition that the image not be destroyed. For example, a landscape containing vivid snow can be added on top of the mountains by labeling the generated image.

### 3.3.1 Generating images with labels

The GAN, as the name suggests, is a generative model in which generator  $G$  attempts to learn a mapping from random noise vector  $z$  to output  $y$  ( $G : z \rightarrow y$ ), where the output can be images or other types of data. However, pix2pix [21] is a model that is based on the conditional GAN, which is a model that learns a mapping from given image  $x$  and random noise vector  $z$  to the output  $y : G : \{x, z\} \rightarrow y$ . To suitably address our question, we replaced the given image  $x$  and random noise vector  $z$  in the concatenation of the border image  $B$  and label map  $Y_{label}$  to transform the mapping into  $G : \{B, Y_{label}\} \rightarrow y$ . The generator  $G$  is trained to generate fake images of such high quality that discriminator  $D$  struggles to distinguish between the fake and real images. This training procedure is illustrated in Fig. 3, and the training steps are listed in Algorithm 2.

### 3.3.2 Objective

The structure of our proposed method, which is based on pix2pix [21]—a variation of conditional GAN—can be expressed as

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D) = & \mathbb{E}_{x,y}[\log D(x, y)] \\ & + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))], \end{aligned} \quad (5)$$

where  $G$  attempts to minimize this objective, but  $D$  tries to maximize it; thus a min–max problem emerges:  $G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D)$ . As previously mentioned, the most critical detail in our method is that the input of the network is the border image  $B$  and the label map  $Y_{label}$  which changes the objective to

$$\begin{aligned} \mathcal{L}_{pixlabel}(G, D) = & \mathbb{E}_{Y_{label}, Y}[\log D(Y_{label}, Y)] \\ & + \mathbb{E}_{Y_{label}, B}[\log(1 - D(Y_{label}, G(C)))], \end{aligned} \quad (6)$$

where  $C$  denotes the concatenation of  $B$  and  $Y_{label}$ .

Research has revealed that if more than one objective is mixed (e.g., L2 distance) [40], the full objective benefits. The training goal of the discriminator remains the same, but the generator's goal is not only to deceive the discriminator but also to approach the ground truth output in the L2 sense. However, it is more suitable to use the L1 distance than L2 because L1 results in less blurring.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]. \quad (7)$$

**Input:**

- 1: N real samples in condition:  $y_1, y_2, \dots, y_n$ .
- 2: The border image of N samples:  $b_1, b_2, \dots, b_n$ .
- 3: The label map of N samples:  $y_{label}^{(1)}, y_{label}^{(2)}, \dots, y_{label}^{(n)}$ .
- 4: Initialize network parameters for Discriminator D, Generator G.
- 5: **for** number of training iterations **do**
- 6:     **for**  $k$  steps **do**
- 7:         Sample minibatch N of real samples  $Y\{y_1, y_2, \dots, y_n\}$  in condition.
- 8:         Sample minibatch N of border images  $B\{b_1, b_2, \dots, b_n\}$ .
- 9:         Concatenate  $Y\{y_1, y_2, \dots, y_n\}$  and  $B\{b_1, b_2, \dots, b_n\}$  into  $C\{c_1, c_2, \dots, c_n\}$ .
- 10:         Update the discriminator  $D$  by ascending its stochastic gradient:

11:

$$\nabla_{\theta_D} \frac{1}{N} \sum_{i=1}^N [\log D(y_i, y_{label}^{(i)}) + \log(1 - D(G(c_i), y_{label}^{(i)}))],$$

$$\text{where } D(y^{(i)}, y_{label}^{(i)}) = True$$

$$\text{and } D(G(c_i), y_{label}^{(i)}) = False \text{ for } i = 1, 2, \dots, N.$$

12:     **end for**13:     **for**  $h$  steps **do**14:         Update the generator  $G$  by descending its stochastic gradient:

15:

$$\nabla_{\theta_G} \frac{1}{N} \sum_{i=1}^N (\log(1 - D(G(c_i), y_{label}^{(i)}))),$$

$$\text{where } y^{(i)} \approx G(c_i) \text{ for } i = 1, 2, \dots, N.$$

16:     **end for**17: **end for**18: Where  $\theta$  represents network parameters, and the gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

---

**Algorithm 2** Image-to-Image translation with labels using minibatch stochastic gradient descent for training, where  $k$  denotes the training steps of discriminator, and  $h$  denotes the training iterations of generator. We use  $k = 1$  and  $h = 3$  in our experiment.

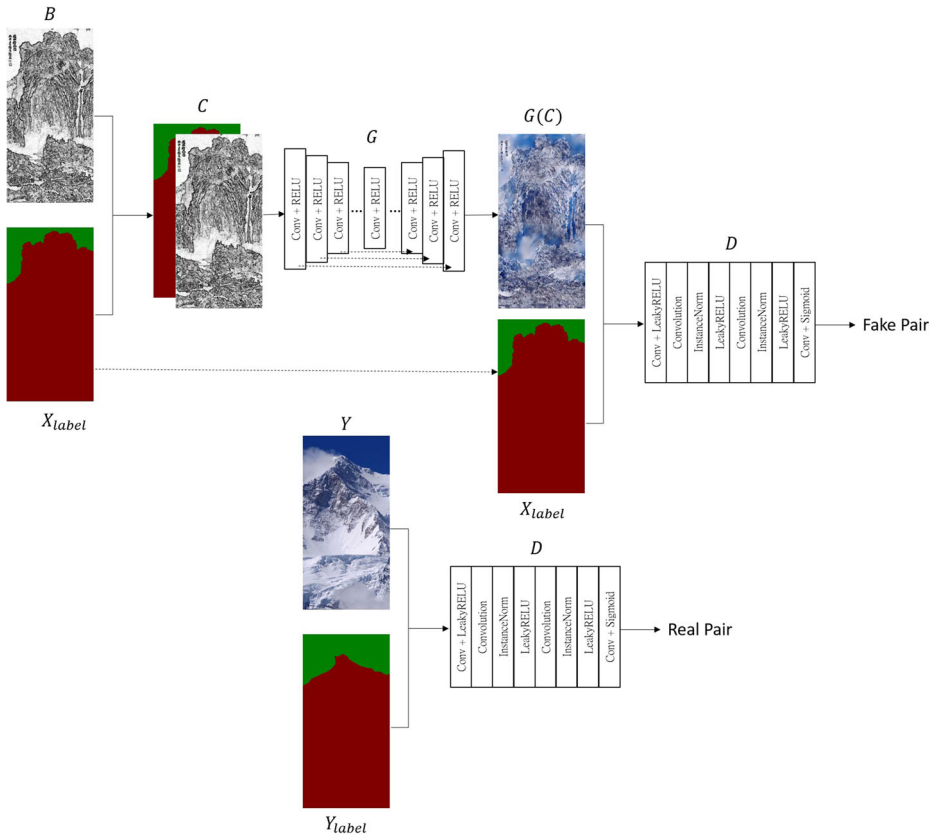
By mixing the two objectives together, we obtain our final objective:

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G), \quad (8)$$

where  $\lambda$  can be adjusted to fit the model. A conditional GAN must be able to produce a high random output if it is to capture the complete entropy of the conditional distribution it models.

### 3.4 BEGAN

In previous sections, we split our proposed method into two parts and separately introduce them in detail in Sections 3.2 and 3.3. In this section, we introduce the full structure, which



**Fig. 3** Fake pair (top) and real pair (below) of the Color part of BEGAN, where  $B$  denotes the border image,  $Y_{label}$  denotes the label map, and  $Y$  denotes the photograph

combines these two parts, and create a new structure named the BEGAN. This structure is called the border enhance GAN because the images we input are not simply transformed into realistic images. Through a specific process, the border can be enhanced to clarify blurry or unclear borders of images transferred from Chinese ink paintings.

### 3.4.1 System architecture

Figure 4 displays the complete structure of the BEGAN. We refer to the first part of our structure, which is used for enhancing the border, as the enhance part. This part visually clarifies the border and separates the different objects. It is crucial for our work because object borders are indistinct in Chinese ink paintings. The second part, which is used for coloring our border image and transforming it into a realistic image, we call the color part. We have replaced the original Pix2Pix input with the border image and object label so that the model generates the target image according to the corresponding edge and label. These two parts of the structure are described in the following sections. These object labels are used not just to train our model but also to stimulate user interaction.

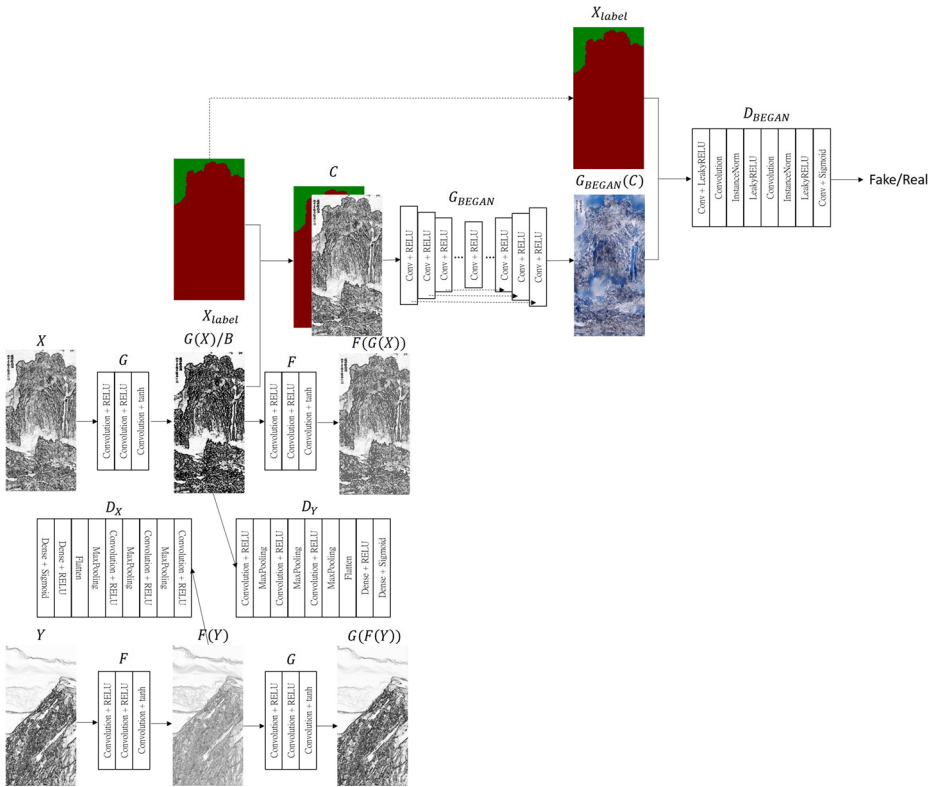


Fig. 4 Structure of BEGAN

### 3.4.2 Network architecture

We use the generator and discriminator structure in [51] for our network architecture. The modules of both the generator and discriminator use the form of convolution-batch normalization-ReLu [20]. The architecture and main features are discussed in the following section.

**The U-net generator with labels** Image-to-image translation involves mapping an input to an output at high resolution. Although the input and output of our problem look different, they have the same underlying structure and share some connections. This means that the structure of the input and output are approximately alike; that is, the information they share can be employed to create links between them. The design of the generator is based on these considerations.

Most related studies [24, 40, 54, 59] for solving this type of problem have employed an encoder–decoder network [19]—a network structure in which the input passes through many layers, and the layers gradually down-sample. A bottleneck layer is located in the middle of the network; after the input passes through the bottleneck, it faces layers that are the reverse of the layers before the bottleneck. With this type of network, the information

progresses from the input through the bottleneck until it reaches the output. In most image translation problems, the goal is to directly transfer information and even share some low-level information between the input and output. For instance, the edges of the input and output in image coloration are consistent, which means they share the same edge-related information.

To obtain a method that bypasses the bottleneck for the generator, we follow the general shape of “U-net” [44] and then added to the network the skip connections between each layer  $i$  and layer  $n - i$ , where  $n$  denotes the total number of layers; thus, layers  $i$  and  $n - i$  concatenate and become a skip connection.

**Using Markovian as discriminator** Numerous related studies have revealed that using L2 loss and L1 loss in image generation problems causes blurry results [28]. In such a scenario, the discriminator only models the high-frequency structure and relies on the L1 term for low-frequency accuracy. For the purpose of simulating high frequencies, we must simply focus on the local patches of images. Thus, the discriminator, called PatchGAN, only has penalties on patches, which means the discriminator attempts to determine whether the image is real or fake in many  $N \times N$  patches and finally averages all feedback to provide the final output of the discriminator. Additionally, a small PatchGAN contains fewer parameters, meaning it can run faster and even be applied to large images. This type of discriminator effectively models the image as a Markov random field, as was reported in [30]. Moreover, the Markov random field is easily obtained in a model of texture [10] and style [9, 17, 29]. Therefore, PatchGAN can be considered a form of texture or style loss.

## 4 Experiments

In this section, we first introduce the characteristics of the system and our data preprocessing procedure. Subsequently, we describe four structures and separately reveal their experimental results, including what aspects they improved, and compare them with existing methods. Furthermore, we quantify the structures to visually represent the effects of the results. Finally, those results are discussed and analyzed, and the advantages and disadvantages of the structures are highlighted. These experimental results enable us to achieve our research goals and produce a suitable structure.

### 4.1 Data description

The main goal of this study was to transform Chinese ink paintings into realistic images. Thus, our training data were numerous Chinese ink paintings and realistic images. To focus on our research goal, we emphasized landscape paintings; thus, the realistic images were photographs containing diverse representations of mountains and the sky. Figures 5 and 6 display the two datasets: the Chinese ink painting dataset and real mountain photograph dataset, respectively. The photograph dataset contained images showing three mountains or mountain groups: Mount Huangshan, Mount Hehuan, and the Himalayas. We obtained the Chinese ink paintings from open data of the National Palace Museum, and we gathered mountain photographs from the Internet.

The Chinese ink paintings presented in Fig. 5 all show different mountains. In terms of the similarities in and problems with these paintings, they were all produced by painters, which means they are unrealistic, lack detail, and have borders not as clear as those in photographs. These problems hamper the direct transformation of the paintings into realistic





**Fig. 5** Chinese ink Painting samples

images; moreover, the viewer may struggle to obtain enjoyment from viewing the transformed images. Mountains from diverse continents were included in the photograph dataset; for example, Tianshan Mountains and Mount Fuji in Asia, the Andes and Rockies in the Americas, Mont Blanc and the Alps in Europe, and Mount Wilhelm and Mount Cook in Oceania. These mountains have unique characteristics and different styles; thus, the user could select the mountains that they were most interested in to finally generate images that met their expectations.

Mount Huangshan appears to consist of rocks and dust, with straight veins and some green trees; these characteristics project a solemn feeling. Mount Hehuan appears green at first, but when it is observed in detail, the large meadow clearly has especially green colors and intricate leaves, and the photograph of this mountain has especially blue sky; a photograph taken in pleasant weather projects a warm feeling. The Himalayas are covered by snow, revealing a cold hue with a navy blue sky; this projects the feeling of extreme cold.

## 4.2 Data preprocessing and selection analysis

In this section, we introduce the data preprocessing steps: cropping the images to the same size, transforming the images into border images, and labelling the images. Subsequently, we explain the method and standard for selecting images to comprise the training dataset. Finally, we analyze the images we selected for the training dataset.



**Fig. 6** From top to bottom, each row contains sample images of Mount Huangshan, Mount Hehuan, and the Himalayas

#### 4.2.1 Image cropping

The images in displayed Figs. 5 and 6 have different shapes. To obtain a common image size and aspect ratio, we had to crop the images to the same shape; once they were a similar shape, they could be fed into our network. After they were cropped into patches, we had many images of size  $256 \times 128$  pixels. A sample of the images is presented in Figs. 7 and 8.



**Fig. 7** The Chinese ink Paintings after cutting into  $512 \times 256$  pixels



**Fig. 8** The real photos after cutting into  $512 \times 256$  pixels

#### 4.2.2 Transformation into border images by using the Sobel method

We prepared our border images by using the Sobel method [15], which obtains the first-order gradient of an image. A sample of the border images is shown in Figs. 9 and 10.

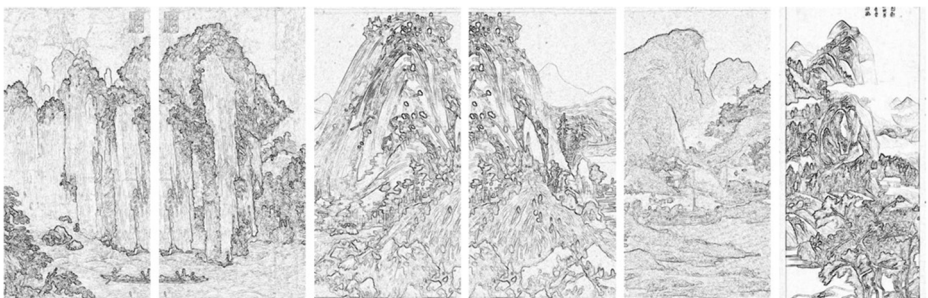
#### 4.2.3 Labelling the images

The last step is to label the images and generate label maps. We used an existing project on GitHub called Labelme [52] to easily label parts of our images and separate the parts. Figures 11 and 12 illustrate the results after our training data were labelled.

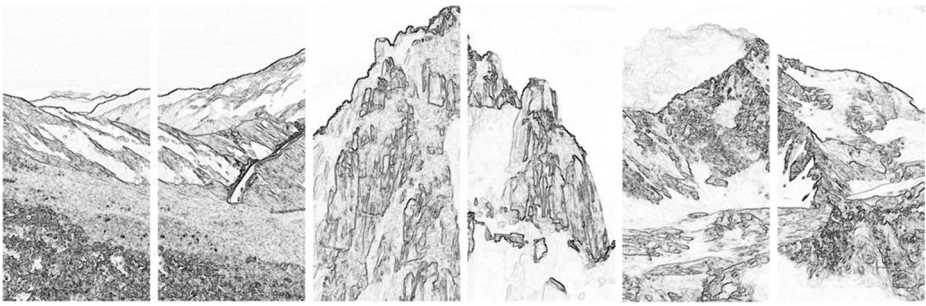
After the images were labelled, they were turned into label maps in which every image was separated into sky and mountain, where 0 denotes the background, 1 denotes the mountain, and 2 denotes the sky. Once the objects in the images had been properly labelled, the generator could generate the appropriate objects.

#### 4.2.4 Data selection analysis

Because we cropped the images in a consistent manner, the images we generated (Figs. 11 and 12) varied considerably, which meant they contained many images that were not helpful to our training and could even have interfered with our training results. To maximize the clarity of our training set, we favored data that were representative and removed images that were deemed irrelevant.



**Fig. 9** Images in Fig. 7 transfer into border images



**Fig. 10** Images in Fig. 8 transfer into border images

For example, some images contained items beyond our main generating target (i.e., sky and mountain) such as words, houses, and people. Thus, we manually removed those redundant images and finally obtained 104 images composed of  $256 \times 128$  pixels and covering all mountain styles in the Chinese ink painting dataset.

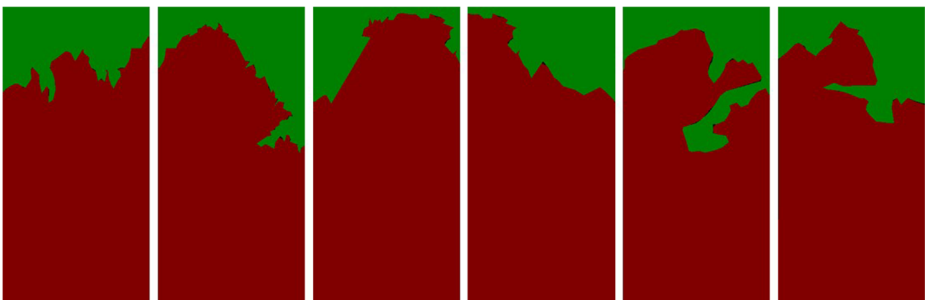
### 4.3 BEGAN experimental results

In this section, we reveal the results of using our BEGAN method to generate realistic images. We conducted four experiments and discuss them separately in the following subsections. Furthermore, we explain in detail the network setup and the main point that was focused on.

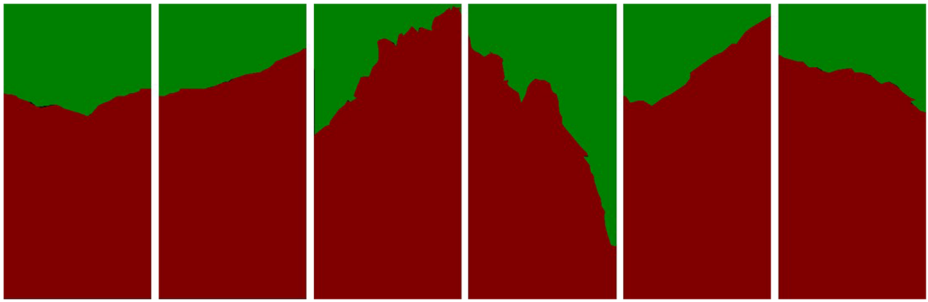
#### 4.3.1 Border enhancement

In this experiment, we implemented the CycleGAN method [60], which is the enhance part of our system, and created a suitable network for our material. The purpose of this step was to enhance the detail of the border image and remove some irrelevant information so that the border image of a Chinese ink painting more closely resembled that of a photograph.

Figure 13 reveals that the border of a Chinese ink painting was similar to the background and also lacked detail; however, after the border was enhanced using the network, the image contained more information. For example, some trees in the image became more lush; in Chinese ink painting, painters only use simple strokes when painting a mountain, but real trees and leaves on mountains of course have more intricate detail. One limitation of our



**Fig. 11** Images in Fig. 7 transfer into label map



**Fig. 12** Images in Fig. 8 transfer into label maps

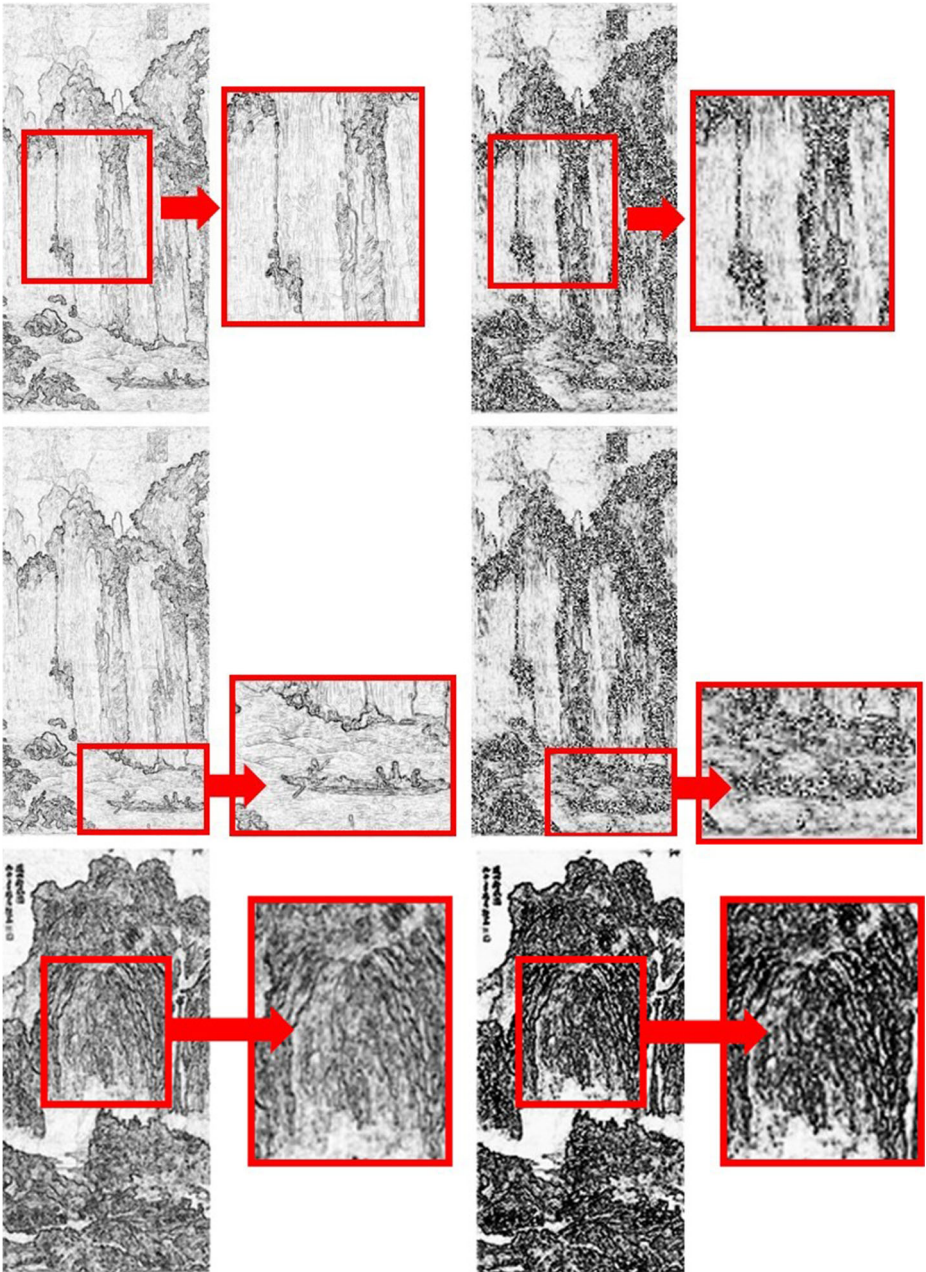
network is that it considers every item that is labelled “mountain” to be a mountain, even objects in an image that are not actually mountains. For example, a boat is depicted in the red boxes in the image in the middle part of Fig. 13, but after the border was enhanced (right part of Fig. 13), the area shows lush underbrush instead of a boat. A large ravine in the red boxes in the lower part of Fig. 13; this ravine also confused our network and generated a fuzzy block of color, which is shown in Fig. 15. More realistic results were obtained when the input images were enhanced, as shown in Fig. 16.

As mentioned previously, because rice paper is used as the canvas in Chinese ink painting, when we used the Sobel algorithm to transform a painting into a border image, some redundant lines remained and affected the generated result. Thus, another goal of the border enhance model was to erase some lines that were not helpful for the image; these lines would negatively affect the results. This problem is not limited to Chinese ink paintings, being evident also for oil paintings. Because the pigment in oil paintings is not as smooth as the color texture in a photograph, the same problem of redundant lines as observed with Chinese ink paintings is discovered in the border image. Thus, we implemented our model with oil paintings to determine whether our model could solve the problem.

In Fig. 14, the red square encloses a piece of the sky; some slight lines are visible in that area, which were caused by the pigment in the oil. After the image was passed through our model, most of the redundant lines in the image were erased; thus, the image more closely resembled a photograph. Through this example, we determined that the border enhance model could mitigate the problem of the border image, including adding some detail and erasing useless lines. Moreover, we discovered that this model is effective not only on Chinese ink paintings but also other types of painting, such as oil paintings.

#### 4.3.2 Border image converted into a realistic image with labels

In this subsection, we reveal some results of the color part, which is the part of our system that concatenates the border image and labels to generate realistic images. As mentioned previously, to focus on the objects we wanted to generate, we not only designated the border image as the input but also concatenated it with the label map; in this manner, the network did not generate the incorrect objects. In conclusion, border enhancement yields two additional effects: (1) clearing useless information and (2) integrating small objects with the background. These effects can help our model generate more realistic images.



**Fig. 13** Red boxes indicate examples of objects that are neither mountain nor sky. Left and right: border images without and with enhancement, respectively

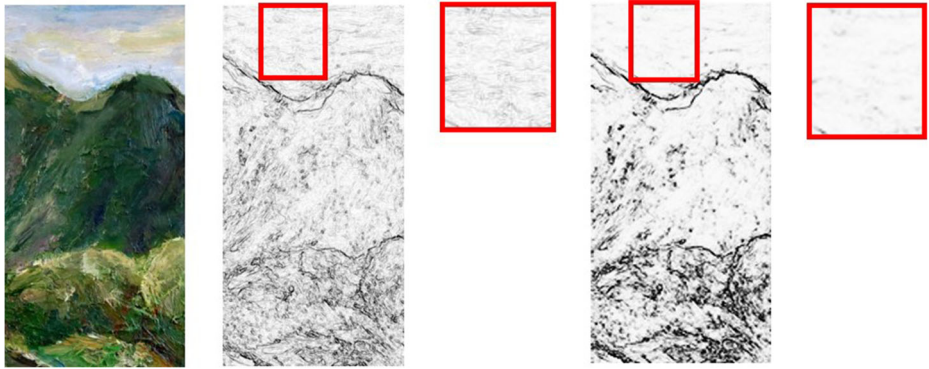


Fig. 14 Result of using the border enhancement method on oil paintings

In Fig. 15, Fan Kuan’s Mountain and Brook Traveler Picture are the experimental images. Because the lines and colors of the original paintings are dark and fuzzy, we used the redrawn and clearer Chinese ink painting as the test object. The main feature of Mount

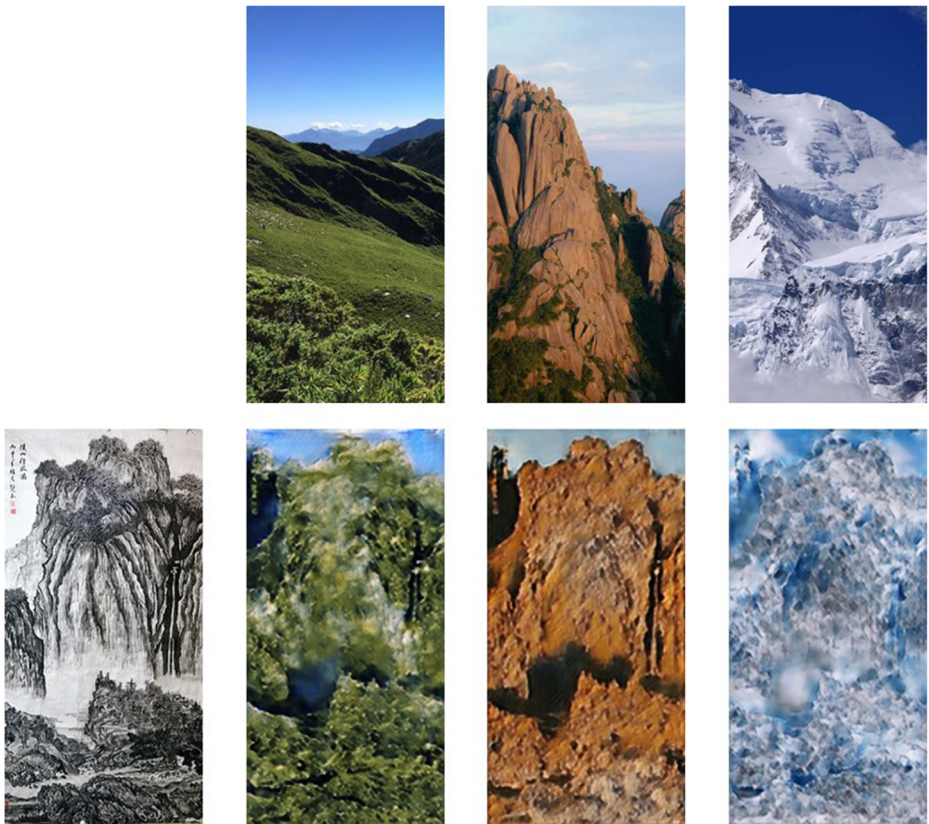


Fig. 15 Result of adding labels when implementing image-to-image translation on Fan Kuan’s Mountain and the Brook Traveler Picture

Hehuan is that it has many green trees and much grass. Someone looking at Mount Hehuan will note that the mountain is green and the texture of the trees is visible. Figure 15 presents the results for the Chinese ink painting in Mount Hehuan style. In addition to retaining the texture of the original painting, the new image has the unique characteristics of this style of mountain as well as color, which is absent in the original painting. For example, in Mount Hehuan style, in addition to the obvious greenness of the whole mountain, the image contains lines resembling trees and leaves on the lines of the mountains; the weather is also favorable, with blue skies highlighting the greenery in the painting. However, khaki is the predominant color on Mount Huangshan; thus, when the Chinese ink painting is transformed into Mount Huangshan style, these mountains appear khaki, and in addition to the change in color, the texture of the mountain is as thick as the soil is. In the final image, we gave the painting Himalayas style. The most obvious characteristic is that the mountain is covered by snow; thus, the painting appears white and cold.

### 4.3.3 BEGAN

The initial results obtained using the Chinese ink paintings were highly colorful; however, they still had poor definition. If look closely at the ridgeline of the valley, the colors are pasted together, which causes the picture to have an unrealistic texture. This was because images generated without border enhancements had unclear details. Thus, we combined the two networks into a new network that not only enhanced the border before transferring the style but also generated a correct and high-definition realistic image.

The results displayed in Fig. 16 are evidently of higher quality than those shown in Fig. 15. The most noticeable difference is in image clarity. After the border was enhanced, the image detail was greater, which led to more realistic and higher definition results. In the image generated using the Mount Hehuan style, trees and leaves were clearly visible, and the shadow was more realistic, making the image brighter and the edges sharper. In the image generated using the Mount Huangshan style, the lines were deeper than they previously were and created the effect of a canyon; moreover, the image resembled Mount Huangshan but in the shape of Fan Kuan's Mountain and the Brook Traveler Picture. In the image using the Himalayas style, the overall color more closely resembles that of the Himalayas, and the white color in the image more closely resembles snow.

In sum, in our experiment, we observed that the border enhancing architecture of our system is of great use. The color part of the system simply generates colored images; however, if the border is enhanced first, images of higher quality and more similar to photographs are obtained.

### 4.3.4 User interaction

In addition to being able to change the style of the mountain, the user can further interact with the image (e.g., modifying and correcting the results), providing a bespoke experience. The advantage of this approach is that the user can make changes until the generated part of the image meets their expectations.

Figure 17 depicts some examples in diverse styles generated through user interaction. The user first selects the main style they desire; subsequently, they identify the area and style they wish to change, resulting in a unique and customized image. The advantage of this approach is that the final result is as unique as each user.





**Fig. 16** Result of BEGAN implemented on Fan Kuan's Mountain and Brook Traveler Picture

#### 4.4 Comparison of existing models

In this section, we discuss samples obtained using existing methods on Chinese ink paintings and discuss relevant problems. We compare two existing methods: (1) image-to-image translation [21] without border enhancement and labelling and (2) style transfer with CNNs [12].

##### 4.4.1 Image-to-image translation

We first compared our model with image-to-image translation. Figure 18 shows the results of image-to-image translation in the Mount Hehuan, Mount Huangshan, and Himalayas styles for Guo Xi's Early Spring.

The image-to-image translation method was also discovered to generate realistic images, but the colors were more turbid and not as vibrant as in our results. In our results when using Mount Hehuan style, the area of the mountain was as green as the trees and the sky was bright and blue; however, when using image-to-image translation (Fig. 18), the image was darker and part of the mountain was not as green as is Hehuan Mount; the sky also was not as cloudy as in the target style. The same problems appeared in both the Mount Huangshan

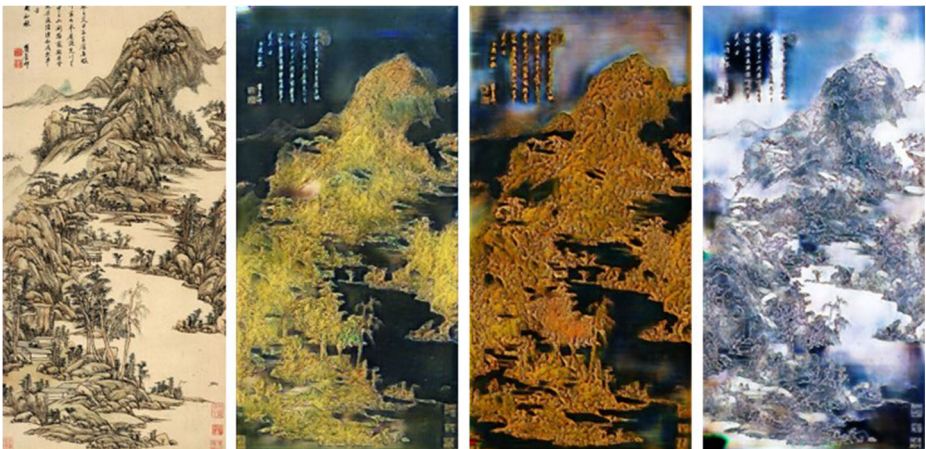


**Fig. 17** User interaction results of mixing multiple styles

and Himalayas styles. The inferior details and color, the most representative characteristic, were the greatest problems.

#### 4.4.2 Style transfer with CNNs

Our system may be regarded as a method of transferring the style of Chinese ink painting; thus, we compared our results with the well-known style transfer method that employs



**Fig. 18** Results of transforming Guo Xi's *Early Spring* into the Mount Hehuan, Mount Huangshan, and Himalayas styles

VGG16 and VGG19. The results are depicted in Fig. 19. This method was used to transform Li Tang's Myriad Ravines with Wind in the Pines into the Mount Hehuan, Mount Huangshan, and Himalayas styles.

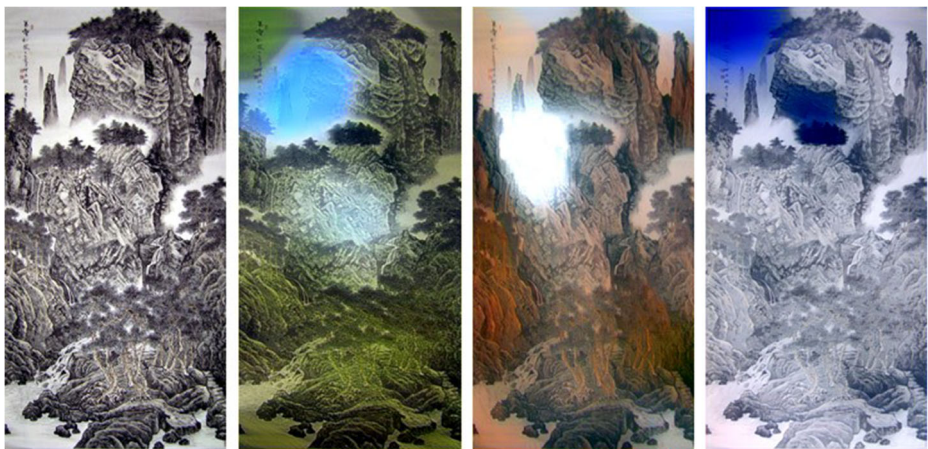
Figure 19 clearly reveals the inferior results. The CNN method only rendered some color in the image and did not achieve our original purpose, which was learning the characteristic style of the mountains and generating an image in that style. Our experimental results revealed that the CNN method is effective for other questions but unsuitable for our question.

#### 4.4.3 Evaluation metrics

In addition to comparing the images we generated, we qualified the results to enable easy evaluation. We used two methods to evaluate performance: the inception score (IS) and Fréchet inception distance (FID). In the following sections, we provide an overview of the two evaluation methods; subsequently, we reveal the metrics for our method, image-to-image translation, and style transfer with CNN.

**IS** The IS [46] is a popular metric for judging the image outputs of GANs. The IS takes a list of images and returns a single floating point number, the score. It is an automatic alternative to having humans grade image quality. The score measures two aspects simultaneously: whether the images have variety and distinctly resemble something. If both conditions are satisfied, the score is high. If either or both are false, the score is low. A higher score is desirable; it means the GAN can generate numerous distinct images. Table 1 shows a comparison of IS for various methods.

However, the IS has some limitations. First, if the system is learning to generate something not present in the training data of the classifier, a low IS is always obtained despite high-quality images being generated because the images are not classified as a distinct class. Second, if images with a different set of labels from the classifier training set are generated, the score may be low. Finally, if the classifier network cannot detect features relevant to the



**Fig. 19** From left to right, the results of transforming Li Tang's Myriad Ravines with Wind in the Pines into the Mount Hehuan, Mount Huangshan, and Himalayas styles

**Table 1** Comparison of Inception Score

Splits	BEGAN		pix2pix		CNN	
	Mean	SD	Mean	SD	Mean	SD
10	1.4838	0.1049	1.5056	0.0815	1.41	0.1614
20	1.4023	0.121	1.4383	0.0964	1.3501	0.1717
30	1.3279	0.1128	1.3611	0.1245	1.2878	0.1601
40	1.2744	0.1182	1.2908	0.1239	1.2255	0.1397
50	1.2342	0.1016	1.2359	0.0966	1.1978	0.1225
60	1.1492	0.1202	1.153	0.1261	1.1428	0.1424
70	1.1033	0.1276	1.1025	0.1239	1.0893	0.1229
80	1.0673	0.1205	1.0641	0.1115	1.0566	0.1052
90	1.0264	0.0709	1.026	0.0716	1.0284	0.0961
100	1.0068	0.0354	1.0068	0.0367	1.0062	0.0348

stated concept of image quality, then poor-quality images may still obtain high scores. For example, people with two heads may be generated, but the system is not penalized for it.

Table 1 reveals that our result was similar to that for pix2pix, and the CNN had obvious low performance. Because our model is based on pix2pix, its evaluation result is similar to that of pix2pix. In the next evaluation matrix, we further evaluate the performance of the BEGAN and pix2pix to confirm the excellent performance of the BEGAN.

**FID** The FID [18] was developed to improve on the IS by actually comparing the statistics of generated samples with those of real samples instead of evaluating generated samples in a vacuum. A lower FID is desirable and corresponds to greater similarity between the real and generated samples as measured by the distance between their activation distributions. Table 2 presents a comparison of the FIDs obtained for the BEGAN and pix2pix results.

The results revealed that although the output was not as superior as the real photographs, BEGAN outperformed the other models in terms of the FID, demonstrating that BEGAN can exhibit superior performance in transforming Chinese paintings and rendering realistic images. The concept of this model can be utilized in other transformation styles for images with blurry borders.

On the basis of the FID, the BEGAN exhibited favorable performance. As previously mentioned, the FID measures the distance between the result and actual photograph; thus, images obtained through the BEGAN more closely resemble photographs. As such, BEGAN mitigates a major problem with pix2pix, making the results more realistic. Many

**Table 2** The comparison of Frechet-Inception-Distance

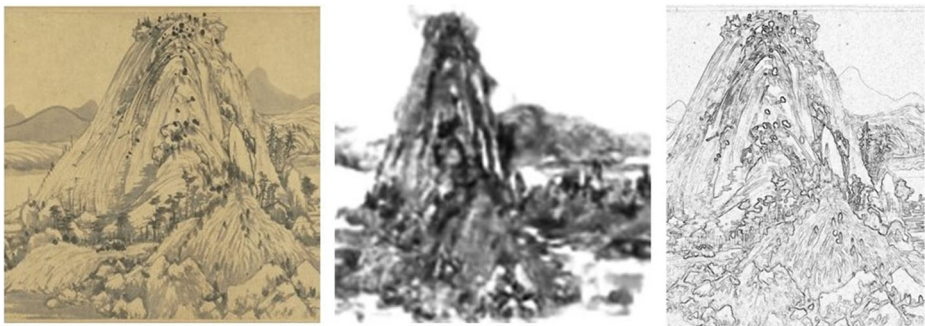
Number of classes	BEGAN	pix2pix
2	242.039	311.142
4	242.039	311.142
8	242.039	311.142
16	247.137	316.648
32	247.137	316.647
64	265.527	325.62

evaluation methods are available; in our case, we not only measured the generated image but also employed an evaluation matrix. Thus, we could conclude that the BEGAN exhibited higher performance than did the methods it was compared with.

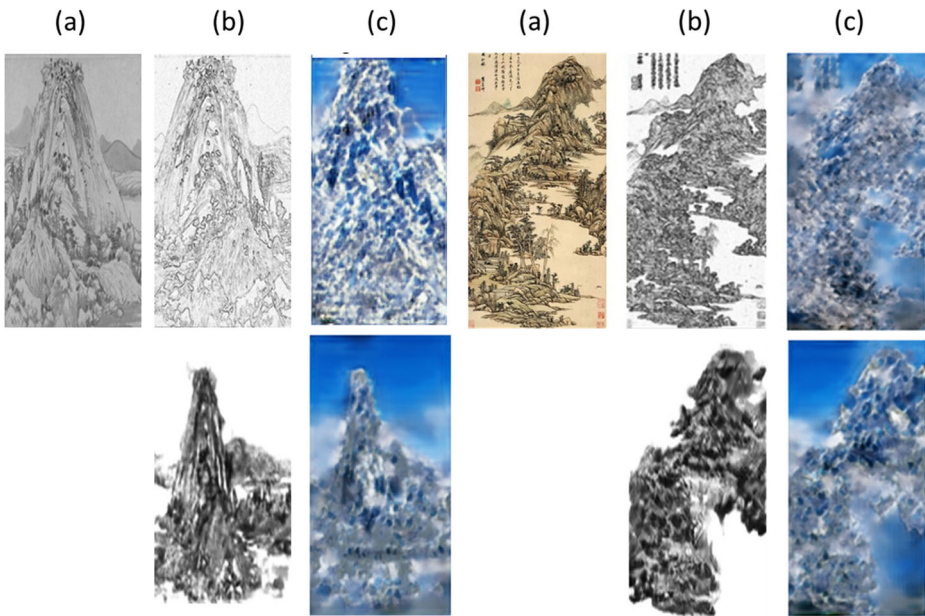
#### 4.4.4 The comparison of proposed method and other benchmark method

In this section, we discuss sketch methods and our proposed method. We present the results obtained using sketch-to-image methods based on Liu et al. [36] and compare it with the original Chinese ink painting (Fig. 20). Sketches and Chinese paintings both lack color and details. In Liu et al. [36], self-supervised learning and an attention module are first used to transform a sketch into a grayscale photograph; their framework is then employed to generate a final realistic image. In our framework, we first obtained border-enhanced images from original images; these images were similar to sketches. Then, we transformed them into realistic photographs by using our border enhance generative adversarial network (BEGAN), which is based on cycleGAN and pix2pix. In the experiment, we compared our method and that of Liu et al. [36] by using the original data and demonstrated that the performance of our method was higher. However, Fig. 20, showing the original image and the images obtained using our proposed method and Liu et al. [36]'s method, indicates that these methods do not result in favorable transformations.

According to the results of the previous paragraph, we believe that our method could capture the object, which is the mountain in our case, well, and finally generate a higher degree of restoration. By using other methods, the shape of the mountain will certainly be reshaped and the boundaries will not be clear. In order to further prove the reliability of our model of the style transferring task, we provide a comparison of the final result between sketch-to-image methods and ours. We first obtain the intermediate outputs of the two models, which are present in the middle section in Fig. 21. Since the brushstrokes of Chinese ink paintings have the characteristic of blurry and unrealistically, our method could depict the edge of the mountain clearly and meticulously. Although the sketch images generated by the sketch-to-images method can capture the full view of mountains, the edge of them is fuzzier and distorted compared with the original paintings. It goes without saying that the result generated by our will emerge more details in the mountain, images presented by the sketch-to-image method is vaguer relatively. All in all, for the task of transforming the Chinese ink painting, BEGAN will produce a better effect in contrast with the sketch-to-image model.



**Fig. 20** Comparison of the sketch-to-image methods. Left to right: original Chinese ink painting, sketch-to-image output, and result obtained using the proposed method



**Fig. 21** Comparison of the intermediate and Himalayas style final results of the sketch-to-image methods. Respectively (a) original Chinese ink painting, (b) sketch output, and (c) the final result using sketch images as input

## 5 Conclusion

In the image-to-image translation of Chinese ink paintings, to address the considerable difference in edge structures between hand-painted and realistic images, we enhanced the details of border images, transformed Chinese ink paintings with scant detail into realistic images, and enabled users to interact with images. Compared with other painting styles transformed using GAN models, Chinese ink paintings do not have distinct boundaries. Therefore, we established BEGAN by combining CycleGAN with pix2pix and adding mountain and sky labels. We are the first to strengthen the borders and landscape details of Chinese ink paintings for image-to-image translation. Although the visual result was acceptable, we observed a considerable difference between the architecture with border enhancement and that without border enhancement; the architecture without border enhancement outperformed that with border enhancement by showing more details in the image and rendered a clearer picture. The object labels are also used to enhance user engagement and are tools enabling user interaction in image generation. The generated results were made consistent with the user's imagination through user involvement in the process; additionally, the user did not simply receive the generated results passively—they were engaged in the creation of Chinese ink paintings. The approach provides users with novel and unique experiences, enabling them to view the value of cultural heritage from new perspectives. Our contributions include the following:

- Enhancing the detail of border maps to turn original Chinese ink paintings lacking detail into real photographs.

- Arranging a labelling mechanism when generating colorful images; the labels are concatenated with the border map to generate accurate and appropriate objects.
- Enabling two-way interaction; users can bring their imagination to life by identifying an area that they would like to change; user input enables limitless variation in the generated images.

Although this research obtained favorable results, some areas may still be improved. In the future, we hope to create a structure that can more effectively add detail to images—for example, some objects on the mountains such as people, houses, and boats, which are essential elements of Chinese ink painting. We would also like to modify some structures in the enhance part to ensure more obvious results of enhancement and create images that more closely resemble photographs. Moreover, we wish to ameliorate the system by shortening the training time and improving the resolution. Finally, we hope to develop a user-friendly and easy-to-understand user interface that provides the user with greater control over the system, enabling more efficient collection of feedback and optimization of the results. With these improvements, we believe that our method will be more mature and practical in the future.

**Acknowledgments** This work was supported in part by the Ministry of Science and Technology, Taiwan, under MOST 111-2622-8-A49-013 -TM1 and MOST 111-2221-E-A49 -125 -MY3; and in part by the Financial Technology (FinTech) Innovation Research Center, National Yang Ming Chiao Tung University.

## Declarations

**Conflict of Interests** All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

## References

1. Chan T-H, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) Pcanet: a simple deep learning baseline for image classification? *IEEE Trans Image Process* 24(12):5017–5032
2. Chen C, Tan X, Wong K-YK (2018) Face sketch synthesis with style transfer using pyramid column feature. In: 2018 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp 485–493
3. Chen L, Wu L, Hu Z, Wang M (2019) Quality-aware unpaired image-to-image translation. *IEEE Trans Multimed* 21(10):2664–2674
4. Chen S (2020) Exploration of artistic creation of chinese ink style painting based on deep learning framework and convolutional neural network model. *Soft Comput* 24(11):7873–7884
5. Cheng Y, Gan Z, Li Y, Liu J, Gao J (2020) Sequential attention gan for interactive image editing. In: Proceedings of the 28th ACM international conference on multimedia, pp 4383–4391
6. Cireřan D, Meier U, Schmidhuber J (2012) Multi-column deep neural networks for image classification. [arXiv:1202.2745](https://arxiv.org/abs/1202.2745)
7. Dai C, Peng C, Chen M (2020) Selective transfer cycle gan for unsupervised person re-identification. *Multimedia Tools and Applications*, 1–17
8. Dou H, Chen C, Hu X, Peng S (2019) Asymmetric cyclegan for unpaired nir-to-rgb face image translation. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1757–1761
9. Efros AA, Freeman WT (2001) Image quilting for texture synthesis and transfer. In: Proceedings of the 28th annual conference on computer graphics and interactive techniques. ACM, pp 341–346
10. Efros AA, Leung TK (1999) Texture synthesis by non-parametric sampling. In: Proceedings of the seventh IEEE international conference on computer vision, vol 2. IEEE, pp 1033–1038

11. Gao W, Li Y, Yin Y, Yang M-H (2020) Fast video multi-style transfer. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3222–3230
12. Gatys LA, Ecker AS, Bethge M (2015) A neural algorithm of artistic style. arXiv:1508.06576
13. Goodfellow I (2016) Nips 2016 Tutorial: generative adversarial networks. arXiv:1701.00160
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
15. Gupta S, Mazumdar SG (2013) Sobel edge detection algorithm. *Int J Comput Sci Manag Res* 2(2):1578–1583
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
17. Hertzmann A, Jacobs CE, Oliver N, Curless B, Salesin DH (2001) Image analogies. In: Proceedings of the 28th annual conference on computer graphics and interactive techniques. ACM, pp 327–340
18. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637
19. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507
20. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167
21. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
22. Jia Z, Yuan B, Wang K, Wu H, Clifford D, Yuan Z, Su H (2020) Lipschitz regularized cyclegan for improving semantic robustness in unpaired image-to-image translation. arXiv:2012.04932
23. Jing Y, Liu X, Ding Y, Wang X, Ding E, Song M, Wen S (2020) Dynamic instance normalization for arbitrary style transfer. In: Proceedings of the AAAI conference on artificial intelligence. vol 34, pp 4369–4376
24. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Springer, pp 694–711
25. Khan A, Ahmad M, Naqvi N, Yousafzai F, Xiao J (2019) Photographic painting style transfer using convolutional neural networks. *Multimed Tools Applic* 78(14):19565–19586
26. Kolkin N, Salavon J, Shakhnarovich G (2019) Style transfer by relaxed optimal transport and self-similarity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10051–10060
27. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
28. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. arXiv:1512.09300
29. Li C, Wand M (2016) Combining markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2479–2486
30. Li C, Wand M (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European conference on computer vision. Springer, pp 702–716
31. Li Y, Tang S, Zhang R, Zhang Y, Li J, Yan S (2019) Asymmetric gan for unpaired image-to-image translation. *IEEE Trans Image Process* 28(12):5881–5896
32. Li Z, Zhou F, Yang L, Li X, Li J (2020) Accelerate neural style transfer with super-resolution. *Multimed Tools Applic* 79(7):4347–4364
33. Liang Y, Lee D, Li Y, Shin B-S (2021) Unpaired medical image colorization using generative adversarial network. *Multimed Tools Applic*, 1–15
34. Lin D, Wang Y, Xu G, Li J, Fu K (2018) Transform a simple sketch to a chinese painting by a multiscale deep neural network. *Algorithms* 11(1):4
35. Liu B, Zhu Y, Song K, Elgammal A (2021) Self-supervised sketch-to-image synthesis. In: Proceedings of the AAAI conference on artificial intelligence. vol 35, pp 2073–2081
36. Liu R, Yu Q, Yu SX (2020) Unsupervised sketch to photo synthesis. In: Computer Vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, pp 36–52
37. Longman R, Ptucha R (2019) Embedded cyclegan for shape-agnostic image-to-image translation. In: 2019 IEEE International conference on image processing (ICIP). IEEE, pp 969–973
38. Lu Y, Wu S, Tai YW, Tang CK, Youtu T (2017) Sketch-to-image generation using deep contextual completion. arXiv:1711.08972




39. Osahor U, Kazemi H, Dabouei A, Nasrabadi N (2020) Quality guided sketch-to-photo image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 820–821
40. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544
41. Peng C, Wang N, Li J, Gao X (2020) Universal face photo-sketch style transfer via multiview domain translation. *IEEE Trans Image Process* 29:8519–8534
42. Peng F, Zhang L, Long M (2018) Fd-gan: face-demorphing generative adversarial network for restoring accomplice's facial image. arXiv:1811.07665
43. Peško M, Trzciński T (2018) Neural comic style transfer: case study. arXiv:1809.01726
44. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
45. Ruder M, Dosovitskiy A, Brox T (2016) Artistic style transfer for videos. In: German conference on pattern recognition. Springer, pp 26–36
46. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2234–2242
47. Shen Y, Luo P, Yan J, Wang X, Tang X (2018) Faceid-gan: learning a symmetry three-player gan for identity-preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 821–830
48. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:abs/1409.1556
49. Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1415–1424
50. Turmukhambetov D, Campbell NeillDF, Goldman DB, Kautz J (2015) Interactive sketch-driven image synthesis. In: Computer graphics forum, vol 34. Wiley Online Library, pp 130–142
51. Tyleček R, Šára R (2013) Spatial pattern templates for recognition of objects with regular structure. In: German conference on pattern recognition. Springer, pp 364–374
52. Wada K (2016) Labelme: image polygonal annotation with Python. <https://github.com/wkentaro/labelme>
53. Wang W, Xu J, Zhang L, Wang Y, Liu J (2020) Consistent video style transfer via compound regularization. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12233–12240
54. Wang X, Gupta A (2016) Generative image modeling using style and structure adversarial networks. In: European conference on computer vision. Springer, pp 318–335
55. Way D-L, Chang W-C, Shih Z-C (2019) Deep learning for anime style transfer. In: Proceedings of the 2019 3rd international conference on advances in image processing, pp 139–143
56. Xue A (2021) End-to-end chinese landscape painting creation using generative adversarial networks. In: Proceedings of the IEEE/CVF Winter conference on applications of computer vision, pp 3863–3871
57. Yao Y, Ren J, Xie X, Liu W, Liu Y-J, Wang J (2019) Attention-aware multi-stroke style transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1467–1475
58. Zhou L, Wang Q-F, Huang K, Lo C-H (2019) An interactive and generative approach for chinese shanshui painting document. In: 2019 International conference on document analysis and recognition (ICDAR). IEEE, pp 819–824
59. Zhou T, Krahenbuhl P, Aubry M, Huang Q, Efros AA (2016) Learning dense correspondence via 3d-guided cycle consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 117–126
60. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Affiliations

Chieh-Yu Chung<sup>1</sup> · Szu-Hao Huang<sup>2</sup> 

Chieh-Yu Chung  
v53828646@gmail.com

<sup>1</sup> Institute of Information Management, National Yang Ming Chiao Tung University,  
Hsinchu 30010, Taiwan

<sup>2</sup> Department of Information Management and Finance, National Yang Ming Chiao Tung University,  
Hsinchu 30010, Taiwan