Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049183

ARTICLE

An Interactive Collaborative Creation System for Shadow Puppets Based on Smooth Generative Adversarial Networks

Cheng Yang[1,2], Miaojia Lou[2,*], Xiaoyu Chen[1,2] and Zixuan Ren[1]

[1]Department of Industrial Design, Hangzhou City University, Hangzhou, 310000, China

[2]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310000, China

*Corresponding Author: Miaojia Lou. Email: mikkayalou@zju.edu.cn

Received: 29 December 2023 Accepted: 28 March 2024 Published: 20 June 2024

ABSTRACT

Chinese shadow puppetry has been recognized as a world intangible cultural heritage. However, it faces substantial challenges in its preservation and advancement due to the intricate and labor-intensive nature of crafting shadow puppets. To ensure the inheritance and development of this cultural heritage, it is imperative to enable traditional art to flourish in the digital era. This paper presents an Interactive Collaborative Creation System for shadow puppets, designed to facilitate the creation of high-quality shadow puppet images with greater ease. The system comprises four key functions: Image contour extraction, intelligent reference recommendation, generation network, and color adjustment, all aimed at assisting users in various aspects of the creative process, including drawing, inspiration, and content generation. Additionally, we propose an enhanced algorithm called Smooth Generative Adversarial Networks (SmoothGAN), which exhibits more stable gradient training and a greater capacity for generating high-resolution shadow puppet images. Furthermore, we have built a new dataset comprising high-quality shadow puppet images to train the shadow puppet generation model. Both qualitative and quantitative experimental results demonstrate that SmoothGAN significantly improves the quality of image generation, while our system efficiently assists users in creating high-quality shadow puppet images, with a SUS scale score of 84.4. This study provides a valuable theoretical and practical reference for the digital creation of shadow puppet art.

KEYWORDS

Shadow puppets; deep learning; image generation; co-create

1 Introduction

Shadow puppetry, a fusion of traditional Chinese arts and crafts with traditional Chinese opera, which has a history of more than 2000 years, was included in the list of Intangible Cultural Heritage of humanity in 2011 [1]. It represents the culmination of human creativity and craftsmanship, embodying significant artistic and cultural value. However, the development and inheritance of shadow puppetry face various difficulties [2]. The old scripts of shadow puppetry do not attract young people, and the art is passed down through face-to-face teaching and incomplete text sources. Additionally, shadow puppetry involves an intricate production process and complicated performance skills. The intricate process of crafting shadow puppets, in particular, can be daunting for modern young

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

individuals, hindering their appreciation of the art's inherent charm [3]. These challenges impede the active participation of ordinary individuals and inhibit the cultivation of widespread interest in shadow puppetry. These aforementioned challenges pose significant barriers to the transmission and continuity of shadow puppetry traditions.

Chinese shadow puppets are diverse and beautiful, created by dedicated practitioners. The development of deep learning and human-computer interaction technology has provided new methods for the inheritance and reactivation of shadow puppetry. Advancements in technology can overcome the limitations associated with traditional methods of shadow puppetry creation, providing innovative approaches for preserving and revitalizing this traditional art form. In terms of shadow puppet image creation: Unique and unprecedented shadow puppet images can be generated by intelligent models. These generation models can learn various styles and design elements to create innovative shadow puppet images. The human-computer collaborative system based on generation models can reduce the difficulty of creating digital shadow puppet characters, enabling users without any expert knowledge to independently create shadow puppet images. In terms of cultural inheritance: The creation system based on generation models, which facilitates user experience quickly with low entry barriers, can promote the interest of young people in traditional shadow puppet art. This aids in the inheritance and development of this culture, allowing traditional art to continue to thrive in the digital age. Therefore, the pursuit of intelligent shadow puppet image creation, leveraging novel technologies to streamline the intricate puppet creation process, not only enhances user experience and reduces barriers to entry but also revitalizes interest in this cherished traditional art form for generations to come.

Previous research predominantly concentrates on shadow puppetry performances [4–6], with less focus on the intelligent creation of shadow puppet images. The intelligent creation of shadow puppetry poses numerous challenges. Using the Component method, Li et al. [3] devised a parameterized template for creating digital shadow puppets, however, this approach hampers users' subjective creativity during the creation process, the solutions are constrained, and the resulting quality is suboptimal. Liang et al. [7] introduced a scene algorithm based on semantic understanding for shadow puppetry scene generation. This method relies on pre-existing shadow puppetry components in the database, limiting the range of solutions. Based on deep learning method, Huang et al. [8] used Cycle-Consistent Generative Adversarial Networks (CycleGAN) to transform facial profile images into shadow puppet head images. However, the generated images lacked quality, and the system could not collaborate with users for creation. While large models like Stable Diffusion [9] can generate shadow puppet images, the limitations of training datasets and the lack of background knowledge within the shadow puppetry domain result in generated from text prompts or contour maps deviating significantly from the traditional style of shadow puppets. Additionally, large model-based creation methods struggle to efficiently collaborate with users, deviating from the production process of traditional shadow puppetry. Existing research highlights three prominent issues: 1) the inability to generate high-resolution images consistent with the style of shadow puppetry, 2) the limitations in assisting users to unleash subjective creativity and create controllable artistic images, and 3) the lack of a high-quality dataset for researching the intelligent generation of shadow puppet images.

In this study, we have developed a human-machine co-creation system for assisting users in creating shadow puppet images while preserving their artistic aesthetic. The system streamlines the traditional shadow puppetry production flow, providing user assistance in three key areas: The drawing process, inspiration stimulation, and content generation. To further enhance the quality of shadow puppet image generation, we propose SmoothGAN, optimized for stable training on high-resolution image tasks, resulting in high-quality images aligned with shadow puppet styles. To train a high-quality model, a dataset of large shadow puppet images is essential. Given the absence of a public

shadow puppet image dataset with uniform backgrounds and pixel sizes, we collected a high-quality dataset comprising 11,211 shadow puppetry head images and 5,256 shadow puppet body images. The research concepts and achievements in this article contribute to the advancement of digital shadow puppetry creation, holding important significance for the inheritance of shadow puppetry.

In summary, this study makes the following contributions:

Firstly, a human-computer co-creation system for shadow puppet images is created, which helps users unleash their creativity to make high-quality shadow puppet images with the assistance of computers in a user-friendly manner. This is research that has never been done before and provides a new idea for the digital protection and development of traditional art.

Secondly, the proposed SmoothGAN model has excellent performance in high-resolution image tasks. The experimental results affirm that our model's generation capability is better than the original model.

Finally, we build a new high-quality shadow puppet images dataset. This dataset encompasses shadow puppetry head images and shadow puppet body images, which may provide a valuable data source for future research in the domain of shadow puppet image generation.

The rest of this paper is organized as follows. Section 2 provides a literature review, introducing traditional Chinese art intelligent generation and image-to-image models. In order to train a high-quality model, a large shadow puppet image dataset is necessary, and Section 3 describes the collection and processing of our shadow puppet dataset. Section 4 illuminates the human-machine system and the SmoothGAN. The evaluation experiments and results are detailed in Section 5. Section 6 discusses the strengths and weaknesses of our system. The conclusions are presented in Section 7.

2 Literature Review

2.1 Intelligent Generation of Traditional Chinese Art Images

In recent years, research on creating traditional Chinese art based on deep learning has become increasingly popularity. Typical traditional Chinese art includes Chinese landscape paintings and Chinese characters.

Zhou et al. [10] introduced an interactive model for generating Chinese landscape paintings based on CycleGAN, allowing users to create artworks by drawing simple lines. Xue et al. [11] proposed an end-to-end Chinese landscape painting generation model, which consists of two-generation networks. One network generates the edge contour map of the landscape painting, while the other transforms this contour map into an authentic landscape painting. Way et al. [12] introduced the Twin Generative Adversarial Network (TwinGan) model, which generates sketches and transforms styles using two sub-networks, proposing a new loss function to preserve the content of the input image. This algorithm can mimic five distinct styles of Chinese landscape painting. Lastly, Wang et al. [13] proposed CCLAP, a controllable landscape painting generation method based on the Latent Diffusion Model. This approach produces Chinese landscape paintings with specified styles using a content generator and a style generator.

Chang et al. [14], using the CycleGAN model, replaced Residual Networks (ResNet) with Densely Connected Convolutional Networks (DenseNet) and added a transfer module for feature extraction. This transformation aimed to convert printed font style into a higher quality and more realistic handwriting style font. Jiang et al. [15] introduced SCFont, a font generation system based on deep stacking networks, consisting of two models. The first model converts the reference font writing trajectory into the target font trajectory, while the second model synthesizes the target font trajectory

into the target font image. Chang et al. [16] proposed a Hierarchical Generative Adversarial Network (HGAN), a method for transforming Chinese fonts. It utilizes a transition network and a hierarchical discriminator to map characters from one font to another while preserving corresponding structural information. This facilitates the conversion of printed fonts into personalized handwritten fonts. Tang et al. [17] proposed Generating Large-scale Chinese Fonts via Recurrent Neural Network (FontRNN) in 2019, breaking down Chinese characters into writing trajectories, training them using Recurrent Neural Network (RNN), and rendering the writing trajectory into target font based on Convolutional Neural Network (CNN). This approach allows computers to generate cursive Chinese character styles. Zhang et al. [18] proposed EMD, a style transfer model addressing challenges faced by traditional style transfer models in handling multiple styles and contents. Its effectiveness and robustness have been verified in the challenges of Chinese character generation. Zhang et al. [19] introduced a Chinese seal carving generation system based on visual knowledge guided and deformation algorithms. This system achieved intelligent generation and layout in Chinese seal carving through the integration of visual knowledge and deformation algorithms, demonstrating superior generation quality compared to the Pix2Pix model.

The above research has realized the style transfer or intelligent generation of traditional artworks based on deep learning, explored methods for computers and users to co-create traditional art and lowered the creation threshold of traditional art. However, the generated artistic images have low pixel resolution and cannot provide users with deeper guidance and collaborative creation. They lack the ability to fully assist users in creating traditional Chinese art images spanning from the initial drawing phase, inspiration cultivation, and content generation while retaining the user's free expression space.

2.2 End-to-End Image Generation

Transforming user-expressed sketches into high-quality shadow puppet images falls within the domain of image-to-image transformation. There are mainly three methods in this field: Style transfer based on neural networks, based on Generative Adversarial Network (GAN) and Diffusion models.

Based on deep neural networks: In 2016, Gatys et al. [20] introduced the style transfer algorithm, which uses convolutional neural networks to extract advanced image features from input images and subsequently recombine their content and style to generate images with diverse artistic styles. Johnso et al. [21] enhanced this algorithm's efficiency by incorporating a perceptual loss function into their neural network training, achieving a remarkable three orders of magnitude speed improvement compared to previous methods. Subsequently, Luan et al. [22] applied style transfer to facilitate realistic transformations in photos. The style transfer algorithm based on deep networks only transfers the color and texture of images but ignores advanced image style semantics.

Based on the Diffusion model: ControlNet [23] serves as a neural network designed for the control of pre-trained model's generation. It extends the functionality of a pre-trained diffusion model to generate new images by incorporating additional input conditions, such as edge mapping images and segmentation mapping images. ControlNet freezes the parameters of Stable Diffusion and zero convolution, achieving end-to-end image generation. However, the Diffusion model requires multiple denoising steps, and although progress has been made in efficient sampling, the generation efficiency is relatively low compared to GAN models, which is difficult to accept in real-time interaction with users.

Based on the GAN: GAN [24] is a deep learning model consisting of two parts: A generator and a discriminator. The generator generates samples that follow the true data distribution through adversarial training. The task of the discriminator is to identify whether the input sample is a pseudo

sample generated by the generator or a real sample. These two parts have opposite objectives and are trained alternately until the samples generated by the generator cannot be distinguished by the discriminator to achieve Nash equilibrium. The joint training of the generator and discriminator involves solving the following maximum and minimum optimization problems:

$$\min_{G}\max_{D} V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

where x represents the image obtained from the real data distribution $p_{data}(x)$, while z represents the latent random extracted drawn from a uniform distribution $p_z(z)$. The specific meanings of abbreviations in the formulas can be found in Table A1 in the Appendix.

Several GANs provide methods for converting sketch input into real images. Pix2Pix [25] is a generic image-to-image translation algorithm based on Conditional GAN (CGAN), which can transform the sketches or color blocks into shoes and street scenes. It uses a combination of adversarial loss and task-related loss. The adversarial loss encourages the generated images to be indistinguishable from real ones, while the task-related loss enforces the desired transformation between domains. However, due to the paired training datasets are often difficult to obtain, Isola et al. [26] presented the CycleGAN in 2017, which can learn the image translation without paired images. It trains two generative models and two discriminative cycle-wise between the input and output. Besides, the cycle consistency loss was designed to ensure the input image and its reconstructed image could be consistent after reverse mapping. Later, Zhu et al. [27] proposed BicycleGAN, which addresses the challenge of generating diverse outputs for a given input. The model simultaneously learns to map from one domain to another (forward mapping) and from the generated domain back to the original domain (backward mapping). This establishes a relationship between latent encoding and target images, allowing the generator to generate different images when given different latent encoding Li et al. [28] proposed a Stacked Cycle-consistent Adversarial Networks (SCAN), which improves the image translation quality by decomposing complex images into multi-level transformations. Additionally, to fully utilize the information from the previous stage, it also designs an adaptive fusion block to learn the dynamic integration between the current stage output and the previous stage output. However, the above studies are unable to train stably in high-resolution image tasks, impacting the model's generation performance.

The SmoothGAN proposed in this study can achieve good performance in high-resolution shadow puppet image generation tasks, and exhibits fast generation efficiency when collaborating with users to create shadow puppet images.

3 Shadow Puppet Images Dataset

There is yet to be a dataset available for generating high-quality shadow puppet images. Therefore, we collected a new dataset of shadow puppet images, which includes 11,211 high-quality shadow puppet head images with unified styles and clear backgrounds, as well as 5,256 shadow puppet body images, to further promote the intelligent generation of traditional Chinese shadow puppetry art. The dataset was established through three main steps.

Collection: We gathered over 30,000 shadow puppet images from search engines and digital museums. To maintain the quality of the dataset, we manually filter out incomplete, low-resolution, and non-traditional shadow puppet art form images.

Cleaning: To achieve superior outcomes in model generation, we preprocess the training data to ensure a consistent background. Firstly, we convert the RGB image into a grayscale image and apply

black-and-white binarization to maximize the differentiation between foreground and background colors. Then use erosion and dilation to identify the maximum and minimum regions in the image. Finally, we replace the background color with white. To ensure the image quality of the expanded, a bicubic interpolation algorithm [29] is utilized to uniformly scale the image and fill it with white pixels, resulting in a resolution of 1024 × 1024 pixels. Fig. 1 shows the cleaning of shadow puppet body images.

Figure 1: Clean the shadow puppet body images to the same white background and pixel size

Obtaining the outer contour of the shadow puppet head images: In order to reduce the difficulty of user expression, we use shadow puppet outer contour image to generate real image. We need to get the dataset pair (outer contour image-real image). The overall contour of the shadow puppetry head image is extracted by the Canny edge detector [30]. We mark the extracted contour set as $C = \{c_1, c_2, \ldots, c_n\}$, and calculate the maximum outer contour through the contour set. The final dataset contains 11211 outer contour images and real images. Fig. 2 shows the processing of shadow puppetry head images.

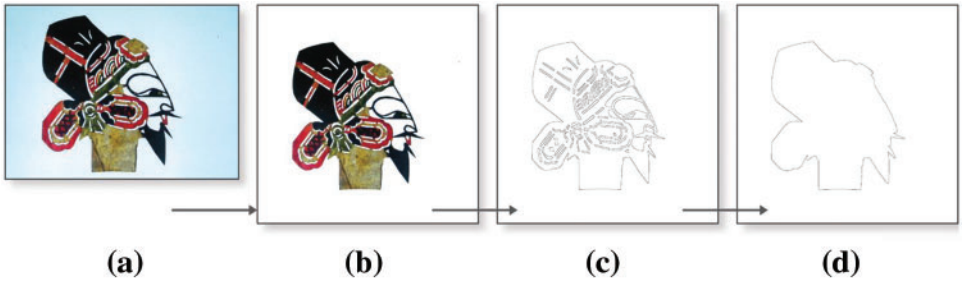

Figure 2: The processing flow of the shadow puppet head images, from left to right, includes the original image (a), the cleaned image (b), the contour image (c) and the outer contour image (d)

4 Methods

4.1 System Design Objectives

The production process of shadow puppets involves several stages, including leather selection, leather making, drawing, copying, engraving, coloring, and splicing. In the drawing stage, there is a "Sample book" that has been handed down through generations, serving as a reference or directly used for copying. Fig. 3a illustrates the "Sample book." During the copying stage, the polished cowhide is placed over the shadow puppet picture, and the lines and patterns of the shadow puppet are depicted on the cowhide with steel needles, then carved, and finally painted and colored. Fig. 3b depicts the copying stage. The shadow puppet characters consist of a shadow puppet head and a shadow puppet body.

(a) (b)

Figure 3: The "Sample book" and the copying stage of producing shadow puppets

In our system, users create shadow puppet images with the assistance of computers and the system flow path follows the traditional handicraft process as much as possible. The system's objectives include three aspects: 1. Reduce the difficulty of drawing; 2. Inspire users; 3. Generate high-quality creative content. To streamline the process and focus on expressive creativity, the steps of leather selection and crafting, which are closely tied to offline activities, are omitted. Key creative steps, such as drawing and coloring, are retained. The system provides reference module to emulate the "sample book". The reference module offers users a foundational understanding of the artistic characteristics of shadow puppetry. During the engraving stage, craftsmen use steel needles to depict the shadow puppet's outer outline on the leather, so the system uses the outer contour image of shadow puppet to serve as input for generating the final image. This approach not only reduces the complexity of drawing but also unleashes the user's creativity. The system can recommend similar sketches based on user-drawn sketches to stimulate user inspiration. The system synthesizes the sketches into high-quality images, presenting them on a shared canvas with the user. At last, user can join their own created shadow puppet head image and satisfactory shadow puppet body image. Fig. 4 illustrates the collaborative creation process of this system. The system is deployed on a website for use.

4.2 System Overview

The framework of our shadow puppet head image creation system is depicted in Fig. 5. The whole system is mainly realized by four key functions: Image contour extraction, intelligent reference recommendation, generation network and color adjustment.

The image contour extraction function can extract the outer contour of the user-selected image from the "sample book" (as illustrated in Fig. 6a), helping the user complete the initial expression quickly during the drawing process. The intelligent recommendation module function stimulates user creativity by recommending eight sketches most similar to the user-drawn sketches, fostering shape and detail thinking. The super-resolution generation network in this study assists users in achieving high-quality shadow puppet head image creation. Finally, the color adjustment function helps users to adjust their results. Fig. 6 shows the main interface of the shadow puppet head creation system, which is divided into the drawing stage and the adjustment stage.

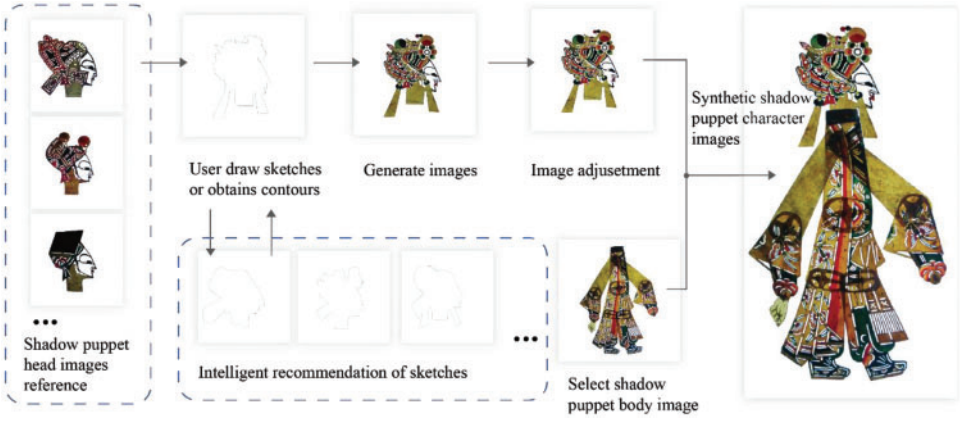

Figure 4: The flow path of collaborative creation of shadow puppet images between the system and users

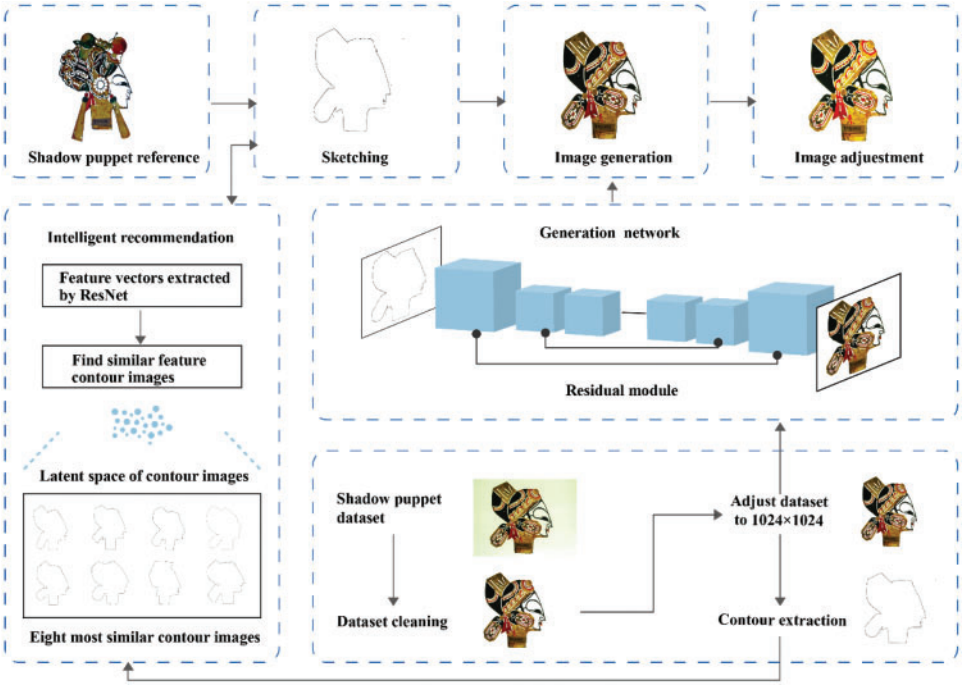

Figure 5: The structure of Shadow puppet head image creation system

Fig. 6a shows the main interface of the drawing stage. This interface consists of four main parts: A drawing tool palette (Area 1), a shared canvas for collaboration between users and computers (Area 2), a "sample book" for reference (Area 3), and an intelligent recommended sketch (Area 4). Area 1 includes a collection of commonly used drawing tools, including brushes, erasers, and the buttons for image generation. After clicking the image generation button, the user proceeds to the adjustment stage (Fig. 6b). Area 2 serves as the main area where users draw sketches and computer-generated images are presented. Area 3 displays the shadow puppet head images that users can refer to. Users can update the reference image by sliding left and right, and click the apply button to extract and display the outer

contour map of the reference image in Area 2. Area 4 presents images recommended by the system based on user sketches. Fig. 6b shows the main interface of the adjustment stage, it consists of two main parts: The control button (Area 1) and the color adjustment tool (Area 2). Area 1 includes buttons for clearing and saving generated images. Area 2 provides users with a tool to adjust the generated image, and users can control the slider to adjust the brightness and saturation of the image.

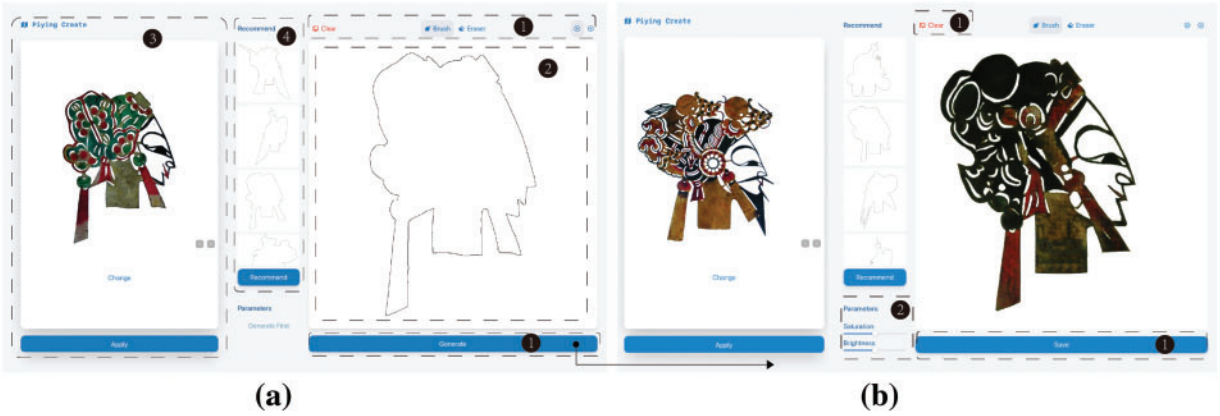

Figure 6: The main interface of the drawing stage (a) and the adjustment stage (b)

4.3 System Implementation

4.3.1 Generation Network Based on SmoothGAN

Our SmoothGAN is based on the Pix2PixHD framework [31] (Henceforth, employ P2P as the representation for Pix2PixHD). P2P achieves higher resolution and more precise image generation compared to Pix2Pix and CycleGAN. It uses three same network structure discriminators (D1, D2, and D3) to discriminate images of different scales. The D used to discriminate images with the smallest resolution, has the largest receptive field, thereby having a more global understanding of the image, and guiding G to generate globally consistent images. The D used to discriminate the image with the largest resolution, is responsible for guiding G to generate the detailed content of the image. Of two generator networks in P2P, G1 generates the global image, and the G2 outputs an image that is twice the width and height of the input in order to be locally enhanced. During training, the residual network G1 is first trained on low-resolution images, and then fine-tuned on high-resolution images together with the residual network G2 to generate images with more realistic details. Fig. 7 shows the overview of the SmoothGAN. The objective function of P2P combines GAN loss $\mathcal{L}_{GAN}$ and feature matching loss $\mathcal{L}_{FM}$ based on multi-scale discriminator. For our task, the objective of G is to translate the contour images into shadow puppet images. While the GAN loss function expression is as follows:

$$\mathcal{L}_{GAN} = \mathbb{E}_{(s,x)}\left[\log D(s,x)\right] + \mathbb{E}_{s}\left[\log(1 - D(s, G(s)))\right] \tag{2}$$

where s represents the contour image, x represents the real image. $G(s)$ represents the fake image generated by the generator.

The feature matching loss is in order to improve the quality of the generated image, which send the generated image and the real image to the discriminator respectively to extract features, and then compute the $L1$ distance of the two features. The loss function expression is as follows:

$$\mathcal{L}_{FM}\left(G, D_k\right) = \mathbb{E}_{(s,x)} \sum_{i=1}^{T} \frac{1}{N_i}\left[\left\| D_k^{(i)}\left(s,x\right) - D_k^{(i)}\left(s, G\left(s\right)\right)\right\|_{L1}\right] \tag{3}$$

where k represents the k-th discriminative network, and i represents the i-th layer within the discriminative networks D_k. s represents the contour image to be converted, x represents the real image, and $G(s)$ represents the target image generated by the generative network. T denotes the total number of layers and N_i denotes the number of elements in each layer of D_k. For the loss $\mathcal{L}_{FM}$, D only serves as a feature extractor. The specific meanings of abbreviations in the formulas can be found in Table A1 in the Appendix.

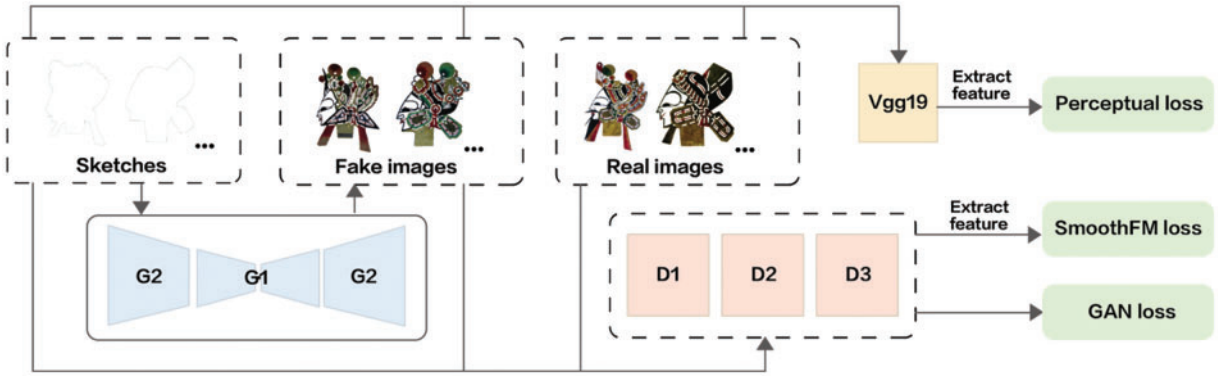

Figure 7: The overview of our SmoothGAN

As the training image pixels in this article are 1024 × 1024. The $\mathcal{L}_{FM}$ is unstable during the training process and is prone to problems such as steepness. To address this issue, in the SmoothGAN model, the L1 loss in $\mathcal{L}_{FM}$ is replaced with the SmoothL1 loss [32]. SmoothL1 improves the problem of zero-point non-smoothness when compared with the L1 loss, and contributes to a more stable and robust training process in super-resolution tasks, enabling the model to converge faster. The expression of the loss function is:

$$\mathcal{L}_{SmoothFM}(G, D_k) = \mathbb{E}_{(s,x)} \sum_{i=1}^{T} \frac{1}{N_i} \left[\left\| D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s)) \right\|_{SmoothL1} \right] \tag{4}$$

Inspired by Güçlütürk et al. [33], we add VGG perceptual loss $\mathcal{L}_{con}$ between generated images and real images to enhance the generative capability of the model. The inclusion of perceptual loss yields improved results in our experiments. The objective function of the SmoothGAN network is:

$$\min_{G} \left(\left(\max_{D_1, D_2, D_3} \sum_{k=1}^{3} \mathcal{L}_{GAN}(G, D_k) \right) + \lambda \left(\sum_{k=1}^{3} \mathcal{L}_{SmoothFM}(G, D_k) + \sum_{k=1}^{3} \mathcal{L}_{con}(G, D_k) \right) \right) \tag{5}$$

where λ controls the weight of the last two items. k represents the k-th discriminative network.

Then joint training of the two generators and three discriminators to achieve Nash equilibrium.

The model in this study was trained on a single GTX 3090 GPU for 200 cycles. The learning rate for the first 100 rounds was set to 0.0002 and gradually decreased from 0.0002 to 0 for the last 100 rounds. The BatchSize was set to 1, the λ was set to 10 and the model training time was approximately 10 days.

4.3.2 Intelligent Reference Recommendation

Since users have a limited basic knowledge of shadow puppets, quickly exploring high-quality shadow puppet outline sketches can be challenging. This system stimulates users' creativity by recommending sketches that resemble their own drawings. Users can receive these recommended

sketches at any time during the drawing process, facilitating the swift expression of their ideas. The system calculates the characteristics of the user's sketches and compares the similarity with the images in the system database, and presents the sketches that are most similar to the user's expression. This assists the users with reference and provides creative inspiration to guide their creation.

To calculate the similarity of image features, we utilize the ResNet-50 pre-trained model [34] for feature extraction. ResNet is constructed from residual blocks, comprising multiple cascaded convolutional layers and a shortcut connection. The shortcut connections enable the network to skip certain layers, computing the sum of the identity map x and the residual map $F(x)$. This architecture ensures information integrity, addressing feature loss issues present in traditional neural networks during information transmission. As a result, it enhances the matching between the sketches recommended by our system and the user's sketches. The network outputs 4096-dimensional feature vectors. We measure the cosine distance between the two 4096-dimensional vectors of the user sketch and the system sketch, defining it as the feature difference. The similarity value range is $[-1, 1]$, where the closer to 1, indicates higher similarity. The expression of cosine similarity is:

$$similarity = \frac{\sum_{i=1}^{n} {v_{user}}^{i} \times {v_{system}}^{i}}{\sqrt{\sum_{i=1}^{n} ({v_{user}}^{i})^2} \times \sqrt{\sum_{i=1}^{n} \left({v_{system}}^{i}\right)^2}} \tag{6}$$

where v_{user} and v_{system} represent the 4096 dimensional vector of the user's hand-drawn sketch and the database sketch, and i represents the component of the vector in the i-th direction.

4.3.3 Color Adjustment

Since the difference in brightness and saturation within the dataset are small, the generated results exhibit relatively high similarity in these two attributes. However, due to aesthetic variations, users may have distinct preferences for the saturation and brightness of the final generated results. To address this, we integrated a color adjustment module into the system, allowing users to make personalized adjustments to the final generated results. This module expands the dimension of human-machine collaborative creation.

To obtain information on the brightness and saturation of the generated image, we convert the image from BGR space to HLS space to obtain these two values. Subsequently, the image is normalized to get the attribute weights of different areas of the image. The value range for brightness and saturation is [0,1], where a smaller value indicates a lower attribute value for the corresponding area in the original image. The formula for controlling saturation or brightness is as follows:

$$Output = (1 + \lambda) \times input \tag{7}$$

where the variable *output* represents the brightness or saturation value of the generated image, while *input* represents the corresponding value of the input image. The weight parameter λ is user-adjustable and lies within the range of $[-1, 1]$.

5 Experiment

5.1 Performance Evaluation Experiments

To evaluate the performance of the proposed SmoothGAN, we compare our method with other end-to-end GAN models, including Pix2Pix and CycleGAN. For these two models, we used the original codes released by the authors, and all parameters used in these models were set to the default settings reported in the papers. Fig. 8 shows the results, which demonstrate that CycleGAN and

Pix2Pix cannot generate a valid shadow puppet head image at a resolution of 1024 × 1024 pixels, while our method generates great results.

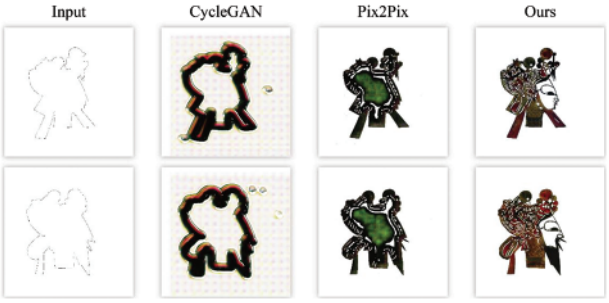

Figure 8: Comparison of our method with other models

An ablation experiment was conducted on two loss functions to demonstrate the effectiveness of the proposed loss function in our work. We trained the P2P, P2P including perceptual loss and P2P using $\mathcal{L}_{SmoothFM}$ on the same dataset with the identical parameters. Fig. 9 illustrates the superiority of the SmoothGAN in generating high-resolution shadow puppet head images, achieving more stable and aesthetically pleasing results. In contrast, images generated by the other three methods had unclear textures in some areas. Subsequently, we invited 20 volunteers to participate in the quantitative evaluation test. Specifically, we employed four models to generate 20 images each. These 80 images were then divided into 20 groups, with each group containing four images generated by the four different models with the same input. Every participant received two groups, and each group was evaluated by two different participants. For each group, participants were asked to score every image on a scale of 1 to 5, with a higher score indicating a greater preference for the image. Fig. 10a shows the results, where our SmoothGAN method received the highest score among the four methods.

Additionally, to comprehensively assess the performance of our proposed method, we employed three commonly used metrics: Fréchet inception distance (FID) [35], Peak Signal-to-Noise Ratio (PSNR), and Structure Similarity Index Measure (SSIM) [36] for quantitative evaluation. SSIM evaluates the structural similarity between the real images and the generated images, while FID and PSNR assess the texture quality of generated images (A lower FID value corresponds to a higher quality of the generated image, whereas higher values of PSNR and SSIM indicate superior image quality). Each of four models generated 11,211 images for calculating these metrics alongside the real images. The comparative results are quantitatively summarized in Fig. 10b. Notably, our method exhibited superior performance, achieving the best scores in FID, RSNR and SSIM.

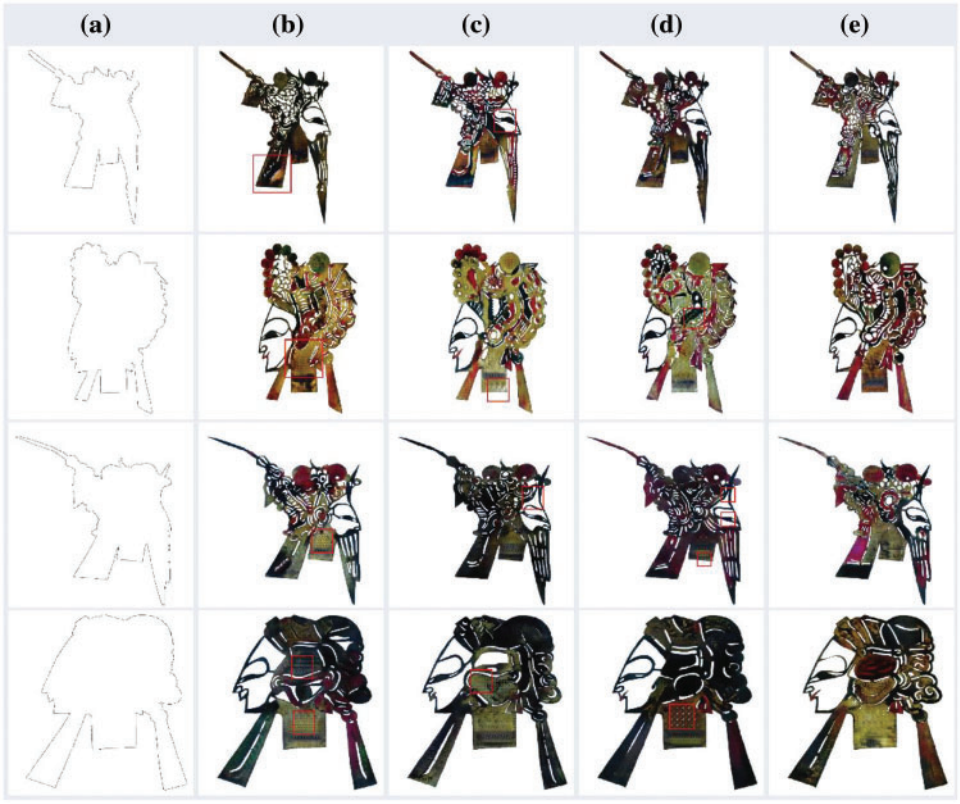

Figure 9: Input sketch (a), P2P (b), P2P using $\mathcal{L}_{SmoothFM}$ (c), P2P including perceptual loss (d), and SmoothGAN (e). The inner rectangle shows the inadequacy of the generated effect

5.2 System Evaluation

The evaluation goals of this system include two: (1) assessing the usability of this system with users; (2) evaluating users' experience with shadow puppet culture after using this system. For the first goal, we use the standard SUS scale and add five new evaluation indicators. These five indicators are derived from the user experience criteria of co-creation between humans and artificial intelligence designed by Oh et al. [37] and the design goals proposed by Benedetii et al. [38], including personalization, fun, flexibility, communication, and kickstart. Each evaluation indicator is designed on a 5-point Likert scale. For the second goal, semi-structured interviews were conducted with the subjects after the experiment. The interviews focused on three themes: (1) the enjoyment and challenges encountered during the experience, (2) the perception of shadow art throughout the experiment, (3) their willingness to further explore and learn about the art of shadow puppetry.

5.2.1 Experimental Method

We recruited 20 participants, encompassing both undergraduate and master's students, with ages ranging from 19 to 29 years. Following a training session on the system's operation, the participants were asked to create a shadow puppet image within 10 min, making a total of 5 images. After completing the images, the participants filled out questionnaires and took part in semi-structured interviews. Some images created by users are shown in Figs. 11 and 12.

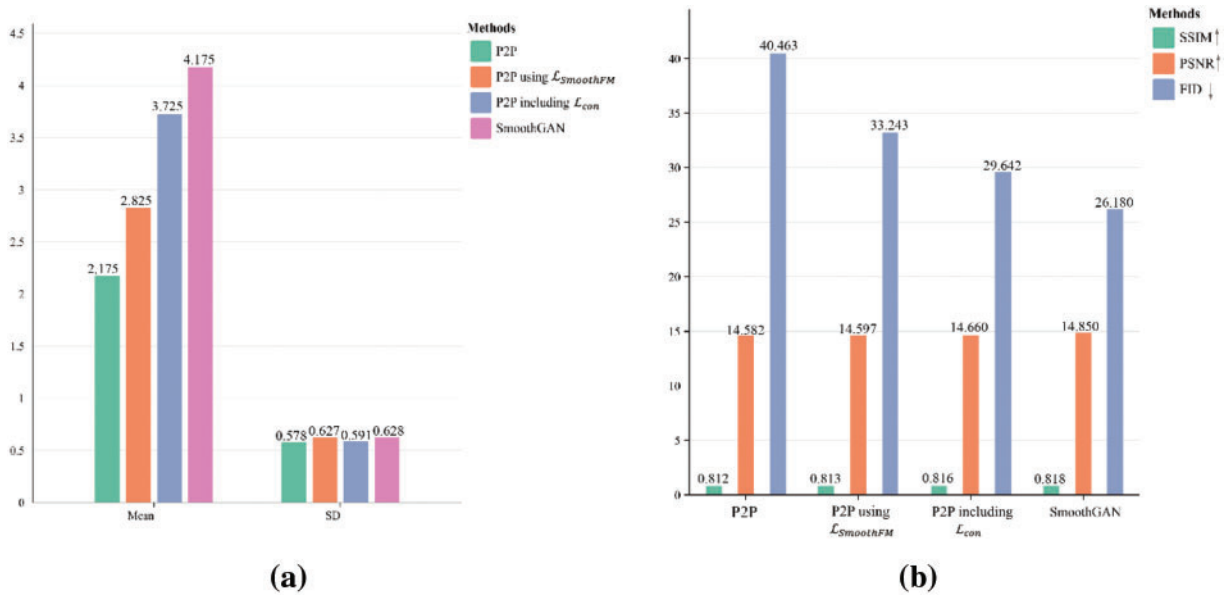

Figure 10: The results of user quantitative evaluation (a) and quantitative evalution (b)

Figure 11: Some shadow puppet head images created by users

5.2.2 Experimental Results

Result 1: Overall, users are satisfied with our system, with an SUS system score of 84.4, as shown in Table 1. The average score of the newly added evaluation indicators exceeds three points, with three indicators (personalization, fun and flexible) exceeding four points, as shown in Table 2. Participants stated that the "sample book" familiarizes them with the artistic characteristics of shadow puppetry in advance. The image contour extraction function helps them to express their creativity quickly, significantly reducing the difficulty of their painting. The recommendation module provides participants with a variety of inspirational sketches, these sketches can stimulate them to think about different contour drawings. Users can quickly apply these recommended sketches to the shared

canvas and modify them. Participants reported that they can create exquisite shadow puppet images without requiring a lot of time for learning. Notably, in the drawing process, users without a painting foundation frequently slide the "sample book" to select a reference picture, extract its outer contour, and proceed with modifications. Users with a painting foundation prefer to directly draw contour maps. Participants described the results as "interesting," "beautiful," and "culturally appropriate." They perceive the system's capacity to cater to the exploration of user personality (mean: 4.2, standard deviation: 0.75).

Figure 12: Some shadow puppet images created by users

Result 2: During the interview, all participants found the drawing process interesting and had expectations for the generated results. Specifically, N5 expressed, "felt a sense of achievement after completing the drawing." while N13 mentioned, "I wasn't originally interested in shadow puppetry, but this way of drawing is really interesting." After the experiment, all participants gained a basic understanding of the color and structure of shadow puppets, suggesting that the system can help users recognize shadow puppets to a certain extent. Nine participants expressed their willingness to spend time appreciating this shadow puppetry culture in their future lives. Additionally, two participants were willing to actively explore this art after experiencing it, indicating that this system can stimulate the interest of ordinary users in this art form. However, some participants expressed dissatisfaction with the system. Two participants expected to generate multiple image selections based on one sketch, and they suggested that the style of the results could be more diverse.

Through experiments, the usability of this system has been substantiated. The system effectively lowers the threshold for creating shadow puppets, enhancing users' comprehension of shadow puppetry, and fostering interest in this traditional culture.

Table 1: Statistical results of SUS scale

Number	Question description	Average score
Q1	I think that I would like to use this system frequently.	4.20
Q2	I found the system unnecessarily complex.	1.50
Q3	I thought the system was easy to use.	4.40
Q4	I think that I would need the support of a technical person to be able to use this system	1.95
Q5	I found the various functions in this system were well integrated.	4.20
Q6	I thought there was too much inconsistency in this system.	1.75
Q7	I would imagine that most people would learn to use this system very quickly.	4.70
Q8	I found the system very cumbersome to use.	1.40
Q9	I felt very confident using the system.	4.30
Q10	I needed to learn a lot of things before I could get going with this system.	1.45

Table 2: Statistical results of five evaluation indicators

Indicators	Personalization	Fun	Flexible	Communicative	Kickstart
Mean	4.20	4.24	3.90	4.35	3.95
SD	0.75	0.89	0.83	0.73	0.80

6 Discussion

In comparison to other methods such as Pix2Pix [25] and CycleGAN [26], the SmoothGAN proposed in this study enhances the quality of image generation. The effectiveness of algorithmic improvement is demonstrated through ablation experiments. Additionally, based on subjective and objective feedback from user experiments, the system exhibits high usability (SUS score of 84.4), which can effectively reduce the difficulty of creating shadow puppet images. In contrast to previous methods based on parameterization [6,7] and deep learning [8], which can effectively help ordinary users create high-quality shadow puppet images from three aspects: User drawing, inspiration stimulation and content generation. The quality of creative results has been significantly enhanced due to algorithmic improvements and the development of high-quality training datasets. This research can effectively aid ordinary users in creating digital shadow puppet images, serving as a significant augmentation to the domain of digital shadow puppet creation.

However, our research also presents several limitations. Firstly, due to the relatively uniform style of the training dataset, the system's generated images lack differentiation in style, failing to meet users' demands for diverse styles. Addressing this issue in the future could involve expanding the dataset and integrating stylization capabilities. Secondly, during the drawing process, some users

without a painting foundation may encounter difficulties such as line jitter and arc deformation when modifying sketches, highlighting the need for system optimization to include a line correction function. Lastly, in the inspiration process, constrained by database limitations and the prioritization of sketches after feature matching, the recommended sketches occasionally diverge significantly from user-drawn sketches. Nevertheless, some users have expressed that these deviated sketches can stimulate them to think about different drawing methods. While the SmoothGAN produces high-quality results, the model's large number of parameters leads to extended training times. This limitation can be addressed in the future by simplifying the model's complexity and incorporating parallel computing capabilities to accelerate training.

7 Conclusion

The intelligent creation of shadow puppetry images will improve the efficiency and quality of shadow puppetry art creation and make shadow puppet images more accessible to people. In this paper, we develop an interactive shadow puppetry creation system that assists users in expressing their creativity from the perspectives of user drawing, inspiration stimulation, and content generation. Furthermore, qualitative and quantitative experimental results demonstrate that the SmoothGAN model proposed in this study exhibits excellent performance in high-resolution image tasks. User research validates the usability of the system, with results indicating its effectiveness in assisting users in creating shadow puppetry images and fostering user's interest in shadow puppetry culture. This study successfully addresses the research gap in interactive shadow puppet creation. Technological innovation and integration will bring new experiences to traditional art. Leveraging deep learning to enable computers to learn the pre-knowledge of traditional arts, user led creative direction, the cooperation between computers and users can improve the efficiency and quality of traditional art creation. This article provides innovative ideas for the protection of traditional art, bearing broad practical significance. Our future work will focus on exploring the transformation of more and higher quality shadow puppet styles, as well as investigating more traditional art forms of human-machine co-creation.

Acknowledgement: All authors extend their gratitude to all the contributors who have collaborated with us on this work.

Funding Statement: This work was supported by the Scientific Research Foundation of Hangzhou City University under Grant No. X-202203 and the Zhejiang Provincial Natural Science Foundation of China under Grant No. LTGY24F030002.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Cheng Yang, Miaojia Lou, Xiaoyu Chen; data collection: Miaojia Lou, Zixuan Ren; analysis and interpretation of results: Cheng Yang, Miaojia Lou; draft manuscript preparation: Cheng Yang, Miaojia Lou, Xiaoyu Chen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this paper are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Lu *et al.*, "ShadowStory: Creative and collaborative digital storytelling inspired by cultural heritage," in *Proc. SIGCHI Conf. Hum. Factor Comput. Syst*, Vancouver, BC, Canada, May 7–12, 2011, pp. 1919–1928. doi: 10.1145/1978942.1979221.

[2] C. Wei, "Research on digital protection and inheritance measures of shaanxi shadow art based on new media times," in *Proc. ICALLH*, 2019, pp. 323–329.

[3] T. Li and W. Cao, "Research on a method of creating digital shadow puppets based on parameterized templates," *Multimed. Tools Appl.*, vol. 80, no. 13, pp. 20403–20422, May 2021. doi: 10.1007/s11042-021-10726-1.

[4] H. Liang, S. Deng, J. Chang, J. J. Zhang, C. Chen and R. Tong, "Semantic framework for interactive animation generation and its application in virtual shadow play performance," *Virtual Real.*, vol. 22, no. 2, pp. 149–165, Jun. 2018. doi: 10.1007/s10055-018-0333-8.

[5] Z. Yan, Z. Jia, Y. Chen, and H. Ding, "The interactive narration of Chinese shadow play," in *Proc.Int. Conf. Virtual Real. Vis. (ICVRV)*, Hangzhou, China, Sep. 2016, pp. 341–345. doi: 10.1109/ICVRV.2016.63.

[6] Y. Shi, F. Ying, X. Chen, Z. Pan, and J. Yu, "Restoration of traditional Chinese shadow play-Piying art from tangible interaction: Transform Chinese shadow play into an interactive system," *Comput. Animat. Virtual Worlds*, vol. 25, no. 1, pp. 33–43, Jan. 2014. doi: 10.1002/cav.1530.

[7] H. Liang, X. Dong, J. Pan, and X. Zheng, "Virtual scene generation promotes shadow puppet art conservation," *Comput. Animat. Virtual Worlds*, vol. 34, no. 5, pp. e2148, Sep. 2023. doi: 10.1002/cav.2148.

[8] X. Huang and J. Huang, "A method for traditional shadow play figure's head regeneration based on generative adversarial network," in *Design Studies and Intelligence Engineering*, vol. 347, pp. 456–463, 2022. doi: 10.3233/FAIA220052.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. Accessed: Dec. 12, 2023. [Online]. Available: http://arxiv.org/abs/2112.10752.

[10] L. Zhou, Q. F. Wang, K. Huang, and C. H. Lo, "An interactive and generative approach for Chinese Shanshui painting document," in *Proc. Int. Conf. Doc. Anal. Recognit. (ICDAR)*, Sydney, Australia, Sep. 2019, pp. 819–824. doi: 10.1109/ICDAR.2019.00136.

[11] A. Xue, "End-to-End Chinese landscape painting creation using generative adversarial networks," 2020. Accessed: Nov. 20, 2021. [Online]. Available: http://arxiv.org/abs/2011.05552

[12] D. L. Way, C. H. Lo, Y. H. Wei, and Z. C. Shih, "TwinGAN: Twin generative adversarial network for chinese landscape painting style transfer," *IEEE Access*, vol. 11, pp. 60844–60852, 2023. doi: 10.1109/ACCESS.2023.3274666.

[13] Z. Wang, J. Zhang, Z. Ji, J. Bai, and S. Shan, "CCLAP: Controllable Chinese landscape painting generation via latent diffusion model," 2023. Accessed: Dec. 14, 2023. [Online]. Available: http://arxiv.org/abs/2304.04156.

[14] B. Chang, Q. Zhang, S. Pan, and L. Meng, "Generating handwritten Chinese characters using CycleGAN," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 12–15, 2018, pp. 199–207. doi: 10.1109/WACV.2018.00028.

[15] Y. Jiang, Z. Lian, Y. Tang, and J. Xiao, "SCFont: Structure-guided Chinese font generation via deep stacked networks," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jul. 2019, pp. 4015–4022. doi: 10.1609/aaai.v33i01.33014015.

[16] J. Chang, "Chinese handwriting imitation with hierarchical generative adversarial network," in *Br. Mach. Vis. Conf. (BMVC)*, Newcastle, UK, Sep. 2018, pp. 290.

[17] S. Tang, Z. Xia, Z. Lian, Y. Tang, and J. Xiao, "FontRNN: Generating large-scale chinese fonts via recurrent neural network," *Comput. Graph. Forum.*, vol. 38, no. 7, pp. 567–577, Oct. 2019. doi: 10.1111/cgf.13861.

[18] Y. Zhang, Y. Zhang, and W. Cai, "Separating style and content for generalized style transfer," in *2018 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8447–8455. doi: 10.1109/CVPR.2018.00881.

[19] K. Zhang *et al.*, "Visual knowledge guided intelligent generation of Chinese seal carving," *Front. Inf. Technol. Electron. Eng.*, vol. 23, no. 10, pp. 1479–1493, Oct. 2022. doi: 10.1631/FITEE.2100094.
[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015. Accessed: Feb. 02, 2023. [Online]. Available: http://arxiv.org/abs/1508.06576.
[21] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016. Accessed: Feb. 02, 2023. [Online]. Available: http://arxiv.org/abs/1603.08155.
[22] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," 2017. Accessed: Feb. 02, 2023. [Online]. Available: http://arxiv.org/abs/1703.07511.
[23] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023. Accessed: Dec. 12, 2023. [Online]. Available: http://arxiv.org/abs/2302.05543.
[24] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020. doi: 10.1145/3422622.
[25] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, IEEE, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.
[26] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2020. Accessed: Oct. 26, 2022. [Online]. Available: http://arxiv.org/abs/1703.10593.
[27] J. Y. Zhu *et al.*, "Toward multimodal image-to-image translation," 2018. Accessed: Nov. 15, 2022. [Online]. Available: http://arxiv.org/abs/1711.11586.
[28] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang and Y. Jiang, "Unsupervised Image-to-Image translation with stacked cycle-consistent adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 08–14, 2018, pp. 184–199.
[29] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981. doi: 10.1109/TASSP.1981.1163711.
[30] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986. doi: 10.1109/TPAMI.1986.4767851.
[31] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," 2018. Accessed: Dec. 14, 2022. [Online]. Available: http://arxiv.org/abs/1711.11585.
[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014. Accessed: Feb. 12, 2023. [Online]. Available: http://arxiv.org/abs/1311.2524.
[33] Y. Güçlütürk, U. Güçlü, R. van Lier, and M. A. J. van Gerven, "Convolutional sketch inversion," *Lecture Note Comput. Sci.*, vol. 9913, no. 8, pp. 810–824, 2016. doi: 10.1007/978-3-319-46604-0_56.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. Accessed: Feb. 13, 2023. [Online]. Available: http://arxiv.org/abs/1512.03385.
[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," 2018. Accessed: Jan. 30, 2024. [Online]. Available: http://arxiv.org/abs/1706.08500.
[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004. doi: 10.1109/TIP.2003.819861.
[37] C. Oh, J. Song, J. Choi, S. Kim, S. Lee and B. Suh, "You help but only with enough details: Understanding user experience of co-creation with artificial intelligence," in *Conf. .Hum. Fact Comput. Syst. Proc. (CHI)*, Montreal, QC, Canada, Apr. 2018, pp. 1–13. doi: 10.1145/3173574.3174223.
[38] L. Benedetti, H. Winnemöller, M. Corsini, and R. Scopigno, "Painting with Bob: Assisted creativity for novices," in *Proc. Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, Honolulu, Hawaii, USA, Oct. 2014, pp. 419–428. doi: 10.1145/2642918.2647415.

Appendix

Table A1: Notations table

Symbol	Description
D	The discriminative network
G	The generative network
s	The contour image
x	The real shadow puppet image
z	Latent random vector
D_k	The k-th discriminative network in SmoothGAN
T	The total number of layers in D_k
N_i	The number of elements in each layer in D_k
$G(s)$	The generated image based on contour image s
$D_k^{(i)}(s, x)$	The features of s and x extracted by the i-th layers of the k-th discriminative network
$D_k^{(i)}(s, G(s))$	The features of s and $G(s)$ extracted by the i-th layers of the k-th discriminative network
$x \sim p_{data}(x)$	A random variable x with distribution $p_{data}(x)$
$z \sim p_{data}(z)$	A random variable z with distribution $p_z(z)$
$\mathcal{L}_{FM}$	The feature matching loss in P2P
$\mathcal{L}_{SmoothFM}$	The smooth feature matching loss in SmoothGAN
$\mathcal{L}_{CON}$	The VGG perceptual loss
$\mathcal{L}_{GAN}$	The Generative Adversarial Network loss
SD	Standard deviation
v_{user}	The user sketch vector extracted by ResNet-50 pre-trained model
v_{System}	The database sketch vector extracted by ResNet-50 pre-trained model