Generating Manga Images from Sketch Via GAN

Ziyuan Liu, Ye Ding, Qi Wan, Zhiyang He, Qing Zhang

Dongguan University of Technology Dongguan, Guangdong, China

Abstract

Image generation has been a popular research topic in computer graphics and computer vision. However, most existing image generation works focus on real-life photographs rather than manga images. Generating manga images directly using photographic image generation methods often results in poor visual performance. We propose a novel manga generation method based on sketch images via GAN. The proposed method is irrelevant to real-life photographs and does not require explicit tags. The resulting manga images are generated based on the sketch images painted by the user. The generated manga image has consistent labels and outlines with the original sketch image and is rendered in manga style. Through our intensive experiments on the public dataset AnimeFace, comparing with the state-of-the-art methods Pix2Pix and SofGAN, the sketch detection model reduces 10.9% FID from SofGAN; and the PSNR of the proposed model is higher than Pix2Pix and SofGAN by 1.4% and 3.9%, respectively. The above qualitative and quantitative evaluations show that our manga generation method has excellent visual performance and serves a controllable and label-free generation of manga images. Statistically, the proposed method outperforms the state-of-the-art.

Keywords

generative adversarial network; image generation; semantic segmentation; style migration

1. INTRODUCTION

Image generation has been a popular research topic of computer graphics and computer vision. Manga-orientated image generation methods often work in two ways: 1) generate manga images randomly through training, such as Pix2Pix [1], Pix2PixHD [2], DCGAN [3], and WGAN [4]. However, it isn't easy to control the results for a random generation model, which is not practical in most application scenarios; and 2) generate manga images based on user-specified tags, such as SofGAN [4], SIS [5]. However, due to the complexity and limited interpretability of computer-generated labels, it is difficult for actual users to specify the desired labels. To conquer the above disadvantages, we propose a novel manga generation method based on sketch images via GAN in this paper.

The proposed manga image generation model is as in Figure 1. The model consists of two parts: 1) the sketch detection model, which generates a feature matrix consisting of feature tags and corresponding positions from the original sketch image through multiple convolution layers. We visualize the feature matrix as a feature map for comparison analytics; and 2) the manga image generator, which takes the feature matrix as input, and generates a corresponding manga image with similar contents to the original sketch image through a texture generator. These two parts are trained separately. We do not need data pairs for training and can generate manga based on random sketches drawn by the user.

We performed a quantitative and qualitative comparison of our generated comic images. The evaluation results show that our system can generate visually pleasing and highly reproducible images that express user's needs.

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control
EMAIL: *dingye@dgut.edu.cn (Ye Ding)

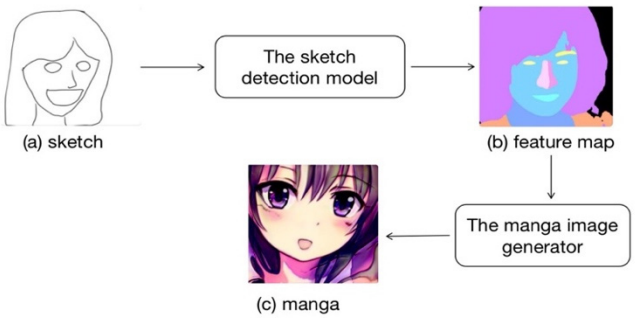

Figure 1. The framework of the sketch-to-manga algorithm

2. METHODS

The framework of the sketch-to-manga algorithm is as in (1). x represents the sketch; $A(x)$ represents the feature matrix; w represents the texture vectors for manga; $G(\cdot)$ represents the transformation of the feature matrix into a manga texture; $S(\cdot)$ represents the whole sketch-to-manga algorithm.

$$S(x) = G(A(x), w). \quad (1)$$

In the following, we will introduce the sketch detection model and the manga image generator in detail.

2.1. The Sketch Detection Model

The function of the sketch feature recognizer is to extract facial features from sketches. Its core focus is to realize the recognition and extraction of key points of sketch faces, which is the key part of realizing the label-free generation of manga.

The module first passes the input image into several convolutional network groups to obtain the feature matrix at different group levels. The higher the level of the feature map, the coarser the detail, while the lower the level, the finer the detail. To enable the module to learn the overall features of the input image while preserving the details of the input image, we take the three layers of high, middle, and low-level feature maps as the input of the subsequent neural network. The high-level feature maps are added and fused with the low level-feature maps and upsampled to obtain new feature maps with more information. Finally, we get three layers from the high, middle, and low layers. These three layers represent feature maps with coarse, medium, and fine features, respectively. These feature maps are passed into a fully convolutional neural network with step size 2 in the ratio of least in the high layer, second in the middle layer, and most in the low layer. So that the number of output layers is continuously reduced during the convolution process to obtain several individual 512-dimensional vectors fed into the generator. Eventually, the sketched facial features are presented in a face with a reasonable facial structure. However, the output obtained at this point is not in the form of a segmentation map and cannot fit the input of the subsequent modules. To extract the feature matrix from the output, we propose the SegExtract part about the real-time semantic segmentation network BiSeNet [6]. It feeds the result into the SegExtract part for segmentation graph extraction and finally gets the feature matrix extracted from the sketch. The network structure of this part is as in (2)(3)(4). x represents the input; out represents the output; cp represents the features that contain contextual semantic information; sp represents the inclusion of spatial information features; $f_i(\cdot)$ represents the 1/i size feature map obtained after downsampling; ARM$(\cdot)$ represents partially optimized ARM [7] features; $FFM(\cdot)$ represents partial feature fusion.

$$\mathrm{sp} = \mathrm{f}_8(\mathrm{x}). \quad (2)$$

$$cp = \mathrm{ARM}(f_{16}(x)) \oplus \left(\mathrm{ARM}\left(\mathrm{f}_{32}(x)\right) \oplus \mathrm{f}_{pooling}\right). \quad (3)$$

$$out = FFM(sp, cp). \quad (4)$$

Since the previously obtained contextual semantic features and spatial features have different output levels and cannot be fused directly, this part reweights them so that the features representing different levels are fused. Therefore, the feature maps obtained from the two parts are finally passed to the FFM

[8] part for feature fusion, and finally, the output layer is convolved and up-sampled to obtain the final output result.

2.2. The Manga Image Generator

To achieve attribute-specific generation, the process of feature matrix to manga images is implemented. We adopt the core idea of StyleGAN [9], a new generative model using an adversarial network progressive resolution enhancement strategy. Starting from a very low resolution, we layer up to a high resolution to control the image attributes. StyleGAN obtains the direction vectors of the specific attributes of the caricature images from a large number of caricature images and then reconstructs the face feature vectors based on the caricature feature direction vectors to generate caricature images with caricature effects consistent with the sketch contours. This process to obtain the style vector. The framework diagram of our caricature image generator model is shown in Figure 2.

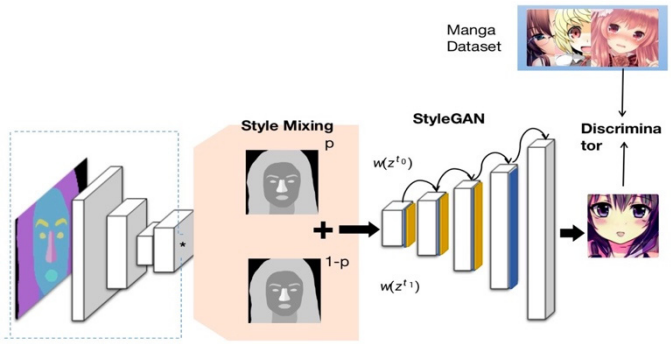

Figure 2. The framework of manga generation

The manga image generator divides the identified feature maps into two regions distance p and $1-p$. We perform random vector sampling of the manga map texture space to get the style vectors z^{t_0} and z^{t_1}, encode and decode the style vector for each label, and fuse the style vectors of the two distance titling maps to get the manga image. Since the goal of StyleGAN is to encode the spatial constraints of the StyleGAN synthesis process while serving the generation quality of the pre-trained StyleGAN, we need to precisely map the encoding conditions to the corresponding parts of the original synthesis process. To achieve this, we formulate the objective function of the training process as in (5). z^{t_0} and z^{t_1} represent the generator mixing two style vectors; $W(\cdot)$ represents the decoding/encoding process of the feature labels; the p's value is between 0 and 1; p represents the similarity between the two styles; β and γ represent the mean and variance of the spatially adaptive normalization parameters; F_{in} represents the matrix of different labels; F_O represents the final generated manga image.

$$F_O = \gamma.(F_{in} * W(z^{t_0}).p + (F_{in} * W(z^{t_1}).(1-p) + \beta. \quad (5)$$

3. Experiments and Results

3.1. Datasets

No sketch dataset has been found to meet the needs of this project from the publicly available sketch dataset resources on the web. The amount of work required to manually draw sketch datasets by hand is too large, and the time cost required is too high. We have used several ways to obtain sketch datasets, as shown in Figure 3.

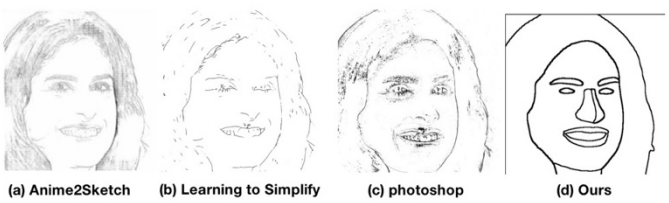

Figure 3. Comparison of sketches generated by different methods

In Figure 3, (a) uses Anime2Sketch [10]; since the sketch generated by (a) contains too much detail, we use a sketch simplification algorithm [11] to generate the sketch to obtain (b); (c) uses the effect of manual processing by Photoshop software; (d) is the effect of processing by our algorithm, and we can see that (d) is more similar to our hand-drawn drawing is more similar.

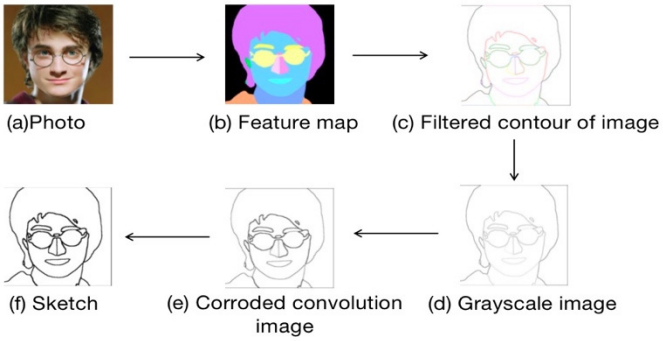

Figure 4. The flow of the sketch production

The process from the real image to the sketch is shown in Figure 4. The process begins by extracting the contours of facial features from real images. The existing semantic segmentation model is used to segment the facial features of the real image to obtain the semantic class markers at the pixel level. Each semantic class is filled with a different distinguishable colour. Figure 5(b) is generated from Figure 5(a). Then the contour filtering operation is done in Figure 5 (b), and the contours of the facial features of the real image are extracted to obtain Figure 5 (c). Then, the extracted contour is greyed out and binarized to make it a binary image to obtain Figure 5 (e). However, the image is not coherent after enlargement, and the contour edges are unlear. Therefore, a corrupted convolution operation is done to obtain the final Figure 5(f), which is the final sketch.

The face dataset used in this paper for sketching is from the publicly available CelebA-HQ, which contains 30,000 high-quality live face images. The manga dataset in our paper is taken from the publicly available manga face dataset AnimeFace downloaded from the face dataset material website (seeprettyface.com).

3.2. The Sketch Detection Model

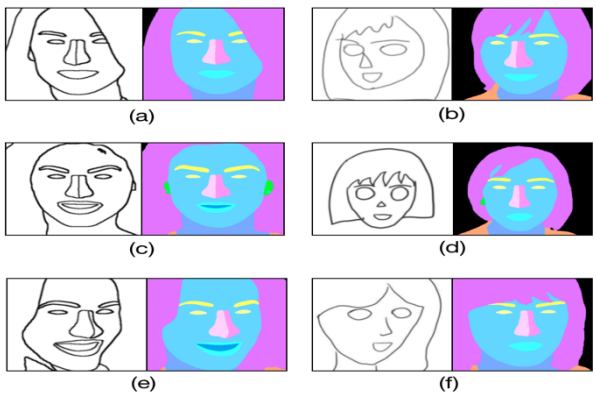

Figure 5. The effect of the sketch detection model

Figure 5(a)-(f) show six visualization effects after recognition by the sketch detection model. The sketches are shown on the left side of the image, and the recognized label images are shown on the right side of the image. Three sketches in (a)(c)(e) are from the sketch dataset, and three sketches in (b)(d)(f) are manually drawn sketches. It can be seen that for the sketches with the same drawing style as the training dataset, the feature maps have extremely high similarity and are consistent with the sketches. For sketches with different styles from the training dataset, the feature maps are consistent with the sketch's facial orientation, five senses layout, and overall structure. In general, the recognition effect of the module is basically as expected.

The comparison of different sketch detection models FID is in Table 1. The smaller the value of FID means, the smaller the distance between images. In the hand-drawn sketch, the FIDs of the segmentation maps generated by the SofGAN model and the real segmentation maps tend to be close to 14. But the FIDs of the sketch detection model are close to 11.96. This comparison confirms that the segmentation maps generated by the sketch detection module implemented in this paper have a small distance from the real segmentation images, and the similarity is high.

TABLE I. FID VALUES OF DIFFERENT MODELS

Modelname	**FID Values**	
	SofGAN	***Ours***
FID	13.42	**11.96**

3.3. The Manga Image Generator

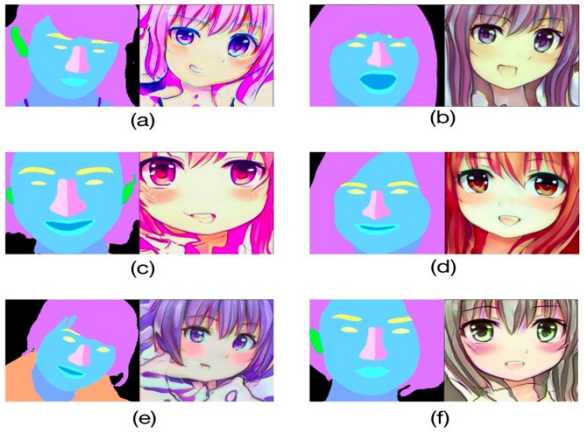

Figure 6. Rendering of different segmentation maps

We use the trained manga image generator model to generate manga avatars for the feature matrix. The experimental results are shown in Figure 6. Figure 6(a)-(f) shows the generation effects of six comic avatars generated from the feature matrix, with the feature matrix on the left and the generated comic avatars on the right of the images. It can be seen that the generated manga avatars correspond to the feature map in terms of facial proportion, facial orientation, distribution of features, and expression display. The model is basically as expected.

3.4. Overall Results

As shown in Figure 7, the system without sketch recognition could not extract labels and generate reasonable comic avatars. The system with sketch recognition performs well on both sketch datasets and manual hand-drawn sketches. It is able to generate comic avatars that correspond to the sketches in terms of facial orientation, layout distribution of the five senses, and expression display, and the generated results are basically as expected.

Figure 8 shows the results of different algorithms for generating comic images. We can see that the outline and pose of our generated cartoon image match better with the hand-drawn cartoon image, which can better express the drawer's intention.

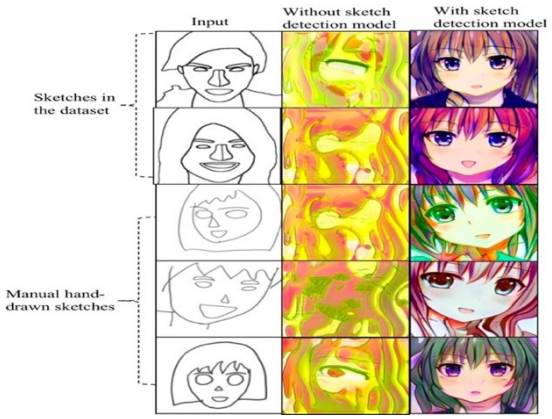

Figure 7. Rendering of different sketches

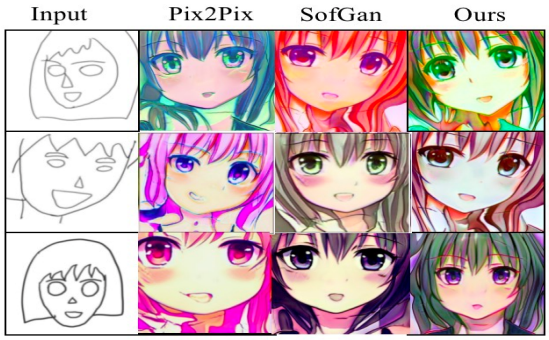

Figure 8. Rendering of different sketch-to-manga algorithms

To objectively confirm the module's effectiveness, we evaluate the similarity between the sketches of this system and the generated comic avatars using an evaluation index like PSNR. The larger the value of PSNR indicates, the smaller the difference value between the two images. From the results in Table 2, we can see that the cartoon generation system implemented in this paper has a better performance compared with other algorithms.

TABLE II. PSNR VALUES OF DIFFERENT MODELS

PSNR Values		
Pix2Pix	***SofGAN***	***Ours***
28.83	**28.14**	**29.24**

4. CONCLUSION AND PROSPECT

We propose an algorithm to generate comic images based on sketches based on social and market demands. We have reviewed the relevant literature published in recent years at home and abroad. We divided the current comic image generation algorithms into three categories: one is to generate comic images with comic features based on real photos, which has the disadvantage that it cannot control the local feature attributes of the images (such as eyes, nose, mouth, etc.) and must have real photos as input; two is to generate comic images randomly based on training data, which has the disadvantage that the generated comic images are random and cannot control the local The third is to generate comic images based on specified labels, which can control the local feature attributes, but the labels must be selected first, and the generated comic images will be strange if the labels are selected incorrectly, and it is tedious to select the labels each time. Based on the shortcomings of the above algorithms, we propose an algorithm that explicitly controls the generation effect of comic images by hand-drawn sketches without selecting labels. The experiments prove that the network model proposed by the author has a high cartoon generation effect on the number of hand-drawn sketches, and the over-fitting problem of the network is reduced after data filtering, which further improves the quality of cartoon image generation. There are

still many unsolved problems in the author's research. For example, the current label recognition performed by the author cannot separate the left and right eyes, and cannot render cartoons for hand-drawn sketches with different postures of the left and right eyes. The future recognition of sketches with more feature details to produce caricature drawings with richer pose contours is an important direction for our research.

5. ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China under grant no. 61976051, U19A2067, and U1811463.

6. REFERENCES

[1] P. Isola, J. Y. Zhu, T. Zhou, and A. E. Alexei, "Image-to-Image Translation with Conditional Adversarial Networks," CVPR, 2017, pp. 1125-1134

[2] 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," Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8798-8807.

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015.

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks". International conference on machine learning. PMLR, 2017: 214-223.

[5] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu, "Semantic Image Synthesis With Spatially-Adaptive Normalization," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2337-2346.

[6] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation," Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[7] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang, "Bi-directional block self-attention for fast and memory-efficient sequence modeling," arXiv preprint arXiv:1804.00857, 2018.

[8] Z. Wu, C. Shen, A. Hengel, "Real-time Semantic Image Segmentation via Spatial Sparsity," arXiv preprint arXiv:1712.00213, 2017.

[9] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, C. Theobalt, "StyleRig: Rigging StyleGAN for 3D Control over Portrait Images," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6142-6151.

[10] X. Xiang, D. Liu, X. Yang, Y. Zhu, and J. P. Allebach, "Adversarial open domain adaption for sketch-to-photo synthesis,".

[11] E. Simo-Serra, S. Iizuka, K. Sasaki, H. Ishikawa, "Learning to simplify: fully convolutional networks for rough sketch cleanup," ACM Transactions on Graphics (TOG), 2016, 35(4): 1-11.