Image Generation of Animation Drawing Robot Based on Knowledge Distillation and Semantic Constraints

Dujuan Wang
School of Information Engineering, Heilongjiang Polytechnic, Heilongjiang 150070, China

***Abstract*—With the development of robot technology, animation drawing robots have gradually appeared in the public eye. Animation drawing robots can generate many types of images, but there are also problems such as poor quality of generated images and long image drawing time. In order to improve the quality of images generated by animation drawing robots, an animation face line drawing generation algorithm based on knowledge distillation was designed to reduce computational complexity through knowledge distillation. To further raise the quality of images generated by robots, the research also designed an unsupervised facial caricature image generation algorithm based on semantic constraints, which uses facial semantic labels to constrain the facial structure of the generated images. The outcomes denote that the max values of the peak signal-to-noise ratio and feature similarity index measurements of the line drawing generation model are 39.45 and 0.7660 respectively, and the mini values are 37.51 and 0.7483 respectively. The average values of the gradient magnitude similarity bias and structural similarity of the loss function used in this model are 0.2041 and 0.8669 respectively. The max and mini values of Fréchet Inception Distance of the face caricature image generation model are 81.60 and 71.32 respectively, and the max and mini time-consuming values are 15.21s and 13.24s respectively. Both the line drawing generation model and the face caricature image generation model have good performance and can provide technical support for the image generation of animation drawing robots.**

***Keywords*—Knowledge distillation; semantic constraints; robot; image; generation**

I. Introduction

A. Background

With the development of technologies such as artificial intelligence, drawing robots have also emerged. As a human-computer interaction task, drawing robots have been applied in many scenarios in life, such as social entertainment. Drawing robots can generate corresponding artistic portraits based on given user photos through algorithms and perform drawing. There are two core issues in drawing robot technology. One is how to use a computer to convert facial photos into high-quality portrait paintings, and the other is how to plan the trajectory of portrait lines so that robots can quickly draw portraits on paper. Current painting robots mainly involve interactive systems and image synthesis algorithms [1-2]. It is very meaningful to draw animations through drawing robots, especially animation images of human faces and portraits, which can reduce the time and labor costs of traditional manual painting. At present, regarding the generation of anime face line drawings, commonly used methods include block-based mechanisms, projection-based methods, generative adversarial learning, and optimization and variants of generative adversarial learning [3]. However, these technologies also have certain shortcomings, resulting in poor image generation quality, long image generation time, and high computational complexity [4].

B. The Method Designed by the Manuscript

With the advancement of deep learning technology, knowledge distillation technology is gradually applied to the compression of different models to reduce the computational complexity of the model [5]. In order to improve the quality of images generated by animation drawing robots, an animation face line drawing generation algorithm based on knowledge distillation was designed, which uses deformable convolution to align features of different scales. The research also designed an unsupervised facial caricature image generation algorithm based on semantic constraints, which uses facial semantic labels to constrain the facial structure of the generated image.

C. The Purpose, Innovation, and Contribution

The research targets to raise the quality of images generated by animation drawing robots from multiple perspectives, reduce the drawing robot's drawing time and operation complexity, and provide good technical support for the wide application of animation drawing robots. The innovation points of the research are mainly reflected in two points. The first point is to combine knowledge distillation, deformable convolution and loss function in the model. The second point is to improve the quality of image generation by drawing robots from the perspectives of anime facial line drawing and facial comic images. The contribution of the research is the improvement of image quality generated by anime drawing robots, the improvement of drawing speed, and the reduction of computational complexity.

D. The Structure of the Manuscript

The research is structured into five sections. Section II is a literature review related to the animation drawing robot image generation. Section III is the specific design of the animation face line drawing generation algorithm and the face caricature image generation algorithm. Results and discussion is given in Section IV and finally, Section V concludes the paper.

II. Literature Review

With the advancement of technologies such as artificial intelligence and robotics, intelligent robots are gradually being utilized to different fields in society. With the development of

the animation industry, more and more researchers have conducted research on image generation for animation drawing robots. Experts such as Ko D K have designed a high update rate method for image generation problems. The method involves low update images, current gripping position and motor current. The research also equipped the robot's gripper with cameras and gripping force sensors. The outcomes denoted that the method designed by this research can generate high update rate images [6]. Liu R and other scholars designed a flexible and robust robot system to solve the problem of autonomous drawing on three-dimensional surfaces, and took two-dimensional drawing strokes and three-dimensional target surfaces as inputs. The system also involves visual recognition, grasping posture reasoning and motion planning. The outcomes denoted that the system is flexible and robust, capable of generating robot motion and successfully drawing three-dimensional strokes [7]. Researchers such as Khanam Z analyzed the impact of gamma radiation on robot vision sensors in nuclear sites by analyzing two images at different dose rates, namely dark images and bright images. Experiment outcomes show that the electrical characteristics change significantly, and when the gamma dose rate is as high as 3Gy/min, the imaging sensor data is unreliable for visual odometry [8]. In order to design a painting robot with style conversion, Wang T and other experts designed a robot-based real-time collaborative drawing system RoboCoDraw. The system involves a generative adversarial network and a random key genetic algorithm. Style transfer is achieved through the generative adversarial network, and path optimization is achieved through the random key genetic algorithm. The results show that the system can generate cartoon face images from real face images [9].

Wu P L and other experts designed an art robot drawing system in order to create pencil sketches. This system can address the issue of pencil wear through tactile sensing function. In addition, this research also uses neural style transfer technology to extract the content and style features of the image, and performs edge detection and further layering on the newly generated image. The results show that the system has good effectiveness in painting and the painting time is less than 30 minutes [10]. In order to allow non-professionals to operate robots as easily as professionals, researchers such as Jens P introduced text-based programming that minimizes robot manufacturing. Furthermore, the drawing of manual instructions on the workpiece before robot machining is investigated. The results show that the method designed by the institute can help non-professionals operate the robot as easily as professionals [11]. Scalera L and other experts conducted drawing experiments to evaluate the performance of the robot architecture, allowing the experimenters to use their eyes to operate the robot's manipulator. Experimental results show that gaze-based human-computer interfaces are beneficial for amputees and patients with various forms of movement disorders [12]. In order to give a brief report on Drawing Fields, Herrmann E W and other scholars explained the use and origin of Drawing Fields. In addition, the report discusses the cultural, ecological and technological resonances of Drawing Fields. The results show that each painting in Drawing Fields corresponds to a different theme [13].

Overall, there is currently massive research related to image generation for animation drawing robots. However, these studies also have certain deficiencies, such as low quality of image generation, single image style, long time-consuming painting, and high computational complexity. In addition, existing methods also have other challenges and limitations, such as inadequate facial feature preservation, incomplete detail texture processing, and high storage space requirements [14-15]. Therefore, to raise the quality of images generated by animation drawing robots, an animation face line drawing generation algorithm based on knowledge distillation was studied and designed, and an unsupervised face comic image generation algorithm based on semantic constraints was also designed. The research targets to raise the quality of images generated by animation drawing robots from multiple perspectives.

III. Design of Facial Portrait Generation Algorithm for Animation Drawing Robots

For the image generation problem of animation drawing robots, the research starts from two directions: animation lines and comic images, and designs a face line drawing generation algorithm based on knowledge distillation and an unsupervised face comic image generation algorithm based on semantic constraints. The study uses knowledge distillation to reduce computational complexity and facial semantic labels to constrain the facial structure of the generated image.

A. Construction of Animation Face Line Drawing Generation Algorithm based on Knowledge Distillation

To raise the quality of the images generated by the animation drawing robot, reduce the drawing time of the image and reduce the complexity of the operation, starting from the face portrait image, two image generation algorithms for animation lines and comics were designed. For the generation of face line images, the research uses knowledge distillation to reduce computational complexity, and uses deformable convolution to align features of different scales. Finally, the study uses boundary loss, style loss and coherence loss to further enhance the quality of line drawings generated by anime drawing robots. The model structure of the line drawing generation algorithm designed by the institute is shown in Fig. 1.

From Fig. 1, the model of the line drawing generation algorithm mainly includes pre-trained teacher network, learning network, distillation loss, input and output. The pre-trained teacher network is a modified model that produces line drawings with better results, and then the study will transfer its intermediate layer knowledge to the student network through knowledge distillation. The network structure used in the study is a two-level nested U-shaped structure to obtain more contextual information. The network structure involves the encoder, decoder and saliency map fusion module, and the U-shaped residual module is involved in the encoder. The structural comparison of the original residual block and the U-shaped residual block is denoted in Fig. 2.

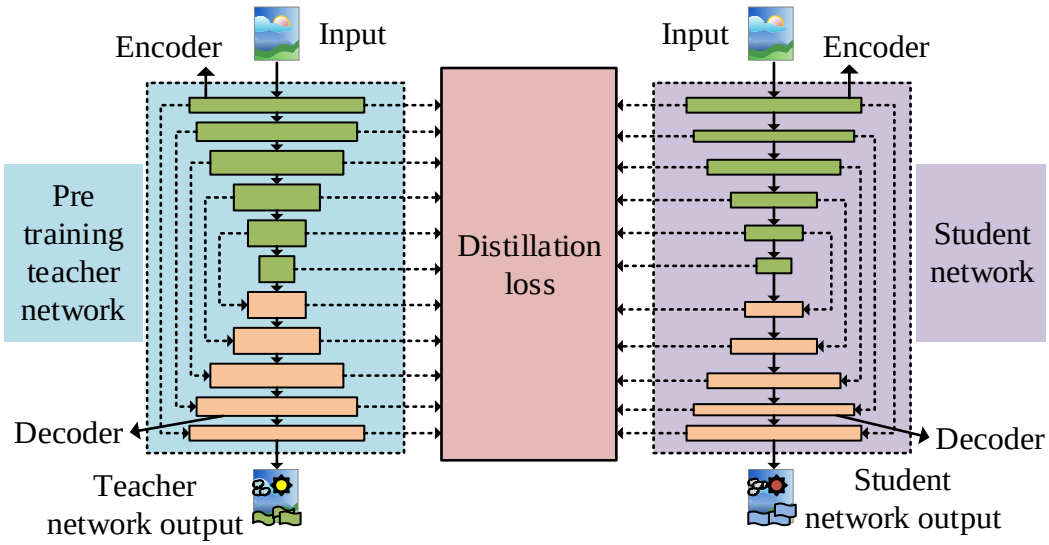


Fig. 1. The model structure of line drawing generation algorithm.

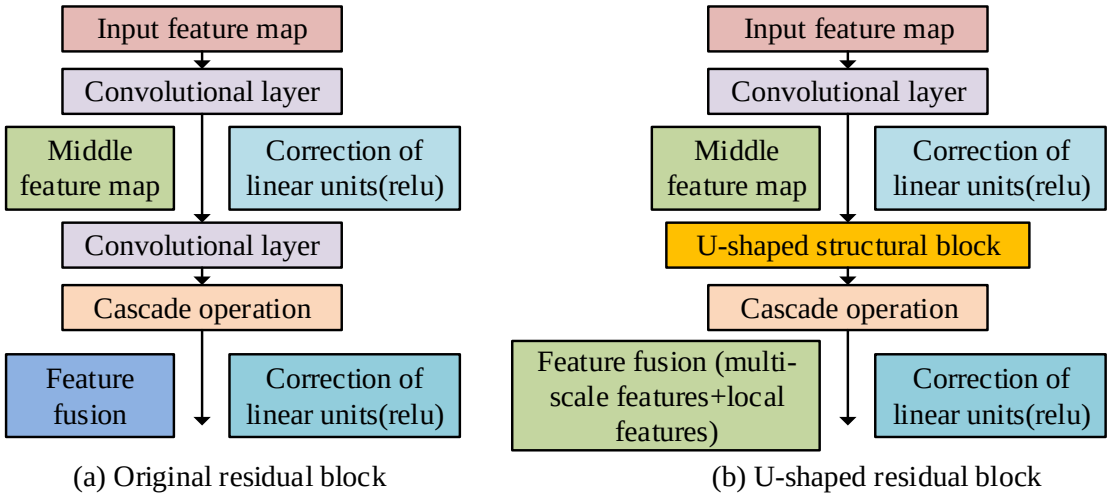


Fig. 2. Structure of U-shaped residual module.

As can be seen from Fig. 2(a), the original residual block mainly includes convolutional layers, modified linear units, intermediate feature maps and feature fusion. From Fig. 2(b), the U-shaped residual block involves convolutional layers, modified linear units, U-shaped structural blocks, multi-scale features and feature fusion. Because the structures of the teacher network and the learning network are both nested U-Nets, in order to avoid damage to target boundary prediction, the research needs to align the upsampling and downsampling features before performing feature fusion. To align features, deformable convolutions were used. The output features at $\hat{a}_p$ any position after convolution a_p are shown in Eq. (1).

$$\hat{a}_p = \sum_{n=1}^{N} \omega_n \times a_p + p_n \quad (1)$$

In Eq. (1), $N = m \times m$, $m \times m$ means the size of the convolution layer and n means the sequence number. ω_n is the weight of the nth convolution sample position, p_n representing the pre-specified offset of the nth convolution sample position. Deformable convolution can adaptively apply to additional offsets at different sample positions, so Eq. (1) can be re-expressed as shown in Eq. (2).

$$\hat{a}_p = \sum_{n=1}^{N} \omega_n \times a_p + p_n + \Delta p_n \quad (2)$$

In Eq. (2), Δp_n represents the additional offset. When deformable convolution is applied to the position information of the down-sampled feature map and the offset field is used as a parameter, the deformable convolution can be aligned by the spatial distance between the position information of the up- and down-sampled feature maps. Therefore, the study selected deformable convolution as the feature alignment function. In order to obtain the trained teacher network, the study introduces boundary loss, style reconstruction loss and coherence loss. In order to further reduce the computational load and model size of the teacher network, the research uses knowledge distillation to

transfer the thinking process and results of the teacher network to the student network, so that students can reach or even exceed the level of the teacher model with a smaller model. To achieve this process, the study adopts feature-based knowledge transfer. The feature-based knowledge distillation loss is shown in Eq. (3) [16].

$$L_{zs}\left(f_t(x), f_s(x)\right) = L_F\left(\Pi_t\left(f_t(x)\right), \Pi_s\left(f_s(x)\right)\right) \quad (3)$$

In Eq. (3), $f_t(x)$ and $f_s(x)$ are the feature maps of the middle layer of the teacher network and student network respectively, $\Pi_t\left(f_t(x)\right)$ and $\Pi_s\left(f_s(x)\right)$ both are conversion functions. $L_F(.)$ represents the distillation loss of matching feature map similarity. The expression of distillation loss is denoted in Eq. (4) [17].

$$L_{dis} = \sum_{h}^{k} kl\left(f_{s_h}, f_{t_h}\right) / d_h \quad (4)$$

In Eq. (4), $kl(.)$ represents the kl divergence function and d_h is the number of channels of the corresponding encoder and decoder. h is the serial number of the channel number. f_{s_h} and f_{t_h} represent the amount of channels of the teacher network and student network, respectively, and k are the number of channels. In addition to boundary loss, style reconstruction loss and coherence loss, the teacher network and student network also involve binary cross-entropy loss and distillation loss, so the loss function of the teacher network is denoted in Eq. (5).

$$L_{teacher} = \beta_1 L_{bce} + \beta_2 L_{style} + \beta_3 L_{boundary} + \beta_4 L_{filter} \quad (5)$$

In Eq. (5), β_1, β_2, β_3 and β_4 are all weight coefficients, L_{bce}, L_{style}, $L_{boundary}$ and L_{filter} are binary cross-entropy loss, style loss, boundary loss and coherence loss respectively. The student network not only needs to use all the loss functions involved in the teacher network, but also needs to use distillation loss. Therefore, the final loss function of the student network is shown in Eq. (6).

$$L_{student} = \beta_1 L_{bce} + \beta_2 L_{style} + \beta_3 L_{boundary} + \beta_4 L_{filter} + \beta_5 L_{dis} \quad (6)$$

In Eq. (6), β_5 it is also the weight coefficient.

B. Design of Unsupervised Face Caricature Generation Algorithm Based on Semantic Constraints

To raise the quality of images generated by animation drawing robots, research has designed an algorithm for generating facial line images. To further raise the quality of images generated by robots, an unsupervised face caricature image generation model based on semantic constraints was designed. The study uses an unsupervised face caricature image generation model to enrich the image style drawn by the robot, and uses group activation mapping and attention modules to avoid the impact of unimportant features on the generated caricature images. In order to better preserve the facial features of human faces, research uses facial semantic labels to constrain the facial structure of the generated images. The network structure of the algorithm in this chapter mainly contains two generators and two discriminators, and both the generator and the discriminator contain attention modules. The specific structure of the face caricature image generation algorithm is shown in Fig. 3.

As can be seen from Fig. 3, the generator mainly includes downsampling, residual block, encoder, auxiliary classifier and group class activation mapping (Group Class Activation Mapping, Group-CAM). The discriminator mainly involves downsampling, encoder, auxiliary classifier and group activation mapping. The face caricature generation algorithm also involves decoders, features, feature weights, face parsing modules and classifiers, where the decoder contains adaptive residual blocks and upsampling. Class activation mapping can retain the spatial information of the image and use it to guide generator training. In addition, class activation mapping can also determine the method of input image categories and enhance the capabilities of the generator and discriminator [18-19]. However, there is a large amount of meaningless data in the saliency map generated by the class activation map in the model, so the study improved it to form the final Group-CAM. The structure of the Group-CAM model is shown in Fig. 4.

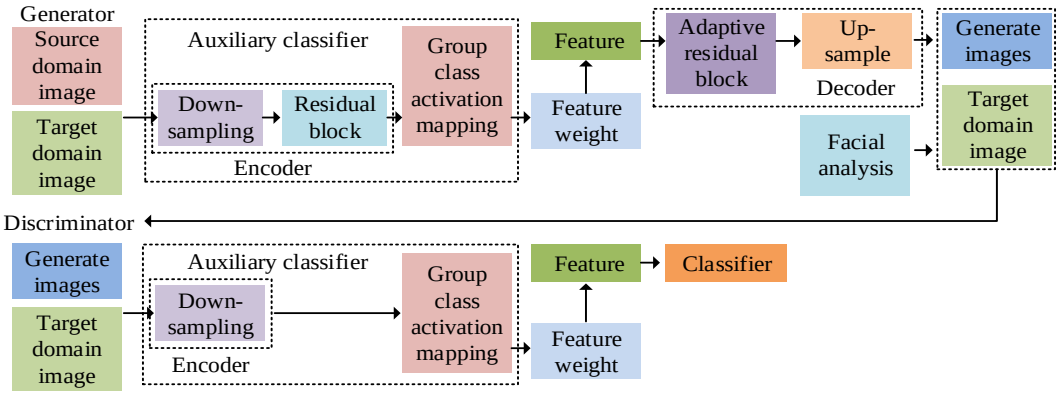


Fig. 3. The specific structure of facial comic generation algorithm.

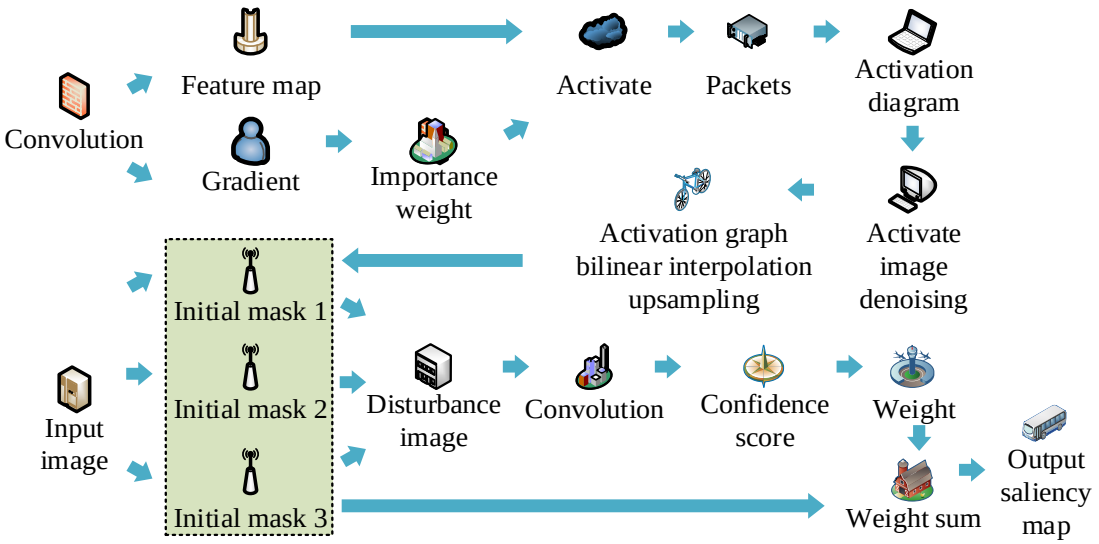


Fig. 4. The structure of the Group-CAM model.

From Fig. 4, the Group-CAM model involves input images, convolutions, feature maps, gradients, importance weights, activations, number of groups, activation maps, activation map denoising, activation map bilinear interpolation upsampling, Initial mask, perturbed image, confidence score, weight sum and saliency map. The initial category mask of the target convolutional layer is shown in Eq. (7).

$$\Upsilon_k^c = \frac{1}{Q}\sum_l\sum_j \frac{\partial F_c(I_0)}{\partial R_{lj}^k(I_0)} \tag{7}$$

In Eq. (7), Q represents the amount of pixels R^k and R^k is the amount of channels of the target layer feature map. $F_c(I_0)$ represents the predicted probability that the input image I_0 is in the class c, l and j the sum is the given number of groups. R_{lj}^k is the sum of the channel numbers of the feature map of the l th group and the j th group of target layers. The initial mask in each group is shown in Eq. (8).

$$M_q = \mathrm{ReLU}\left(\sum_{k=q\times g}^{(q+1)\times g-1} \Upsilon_k^c R^k\right) \tag{8}$$

In Eq. (8), $q \in \{0,1,\ldots,G-1\}$, G denotes the amount of groups of all feature maps and their corresponding importance weights. g denotes the amount of feature maps in each group. Since the initial mask is visually noisy, the study uses a denoising function to process it, and scales the value of the initial mask to [0, 1] through normalization. The initial mask processing process is shown in Eq. (9).

$$M'_q = \frac{M_q - \min(M_q)}{\max(M_q) - \min(M_q)} \tag{9}$$

In Eq. (9), M_q' represents the smoother mask generated by the activation map, $\min(M_q)$ and $\max(M_q)$ are the mini and max values of the initial mask, respectively. Afterwards, the study uses bilinear interpolation for upsampling. When generating saliency maps, blur operations are required. The calculation of the blurred image is shown in Eq. (10) [20].

$$I'_q = I_0 \odot M'_q + \tilde{I}_0 \odot (1 - M'_q) \tag{10}$$

In Eq. (10), $\tilde{I}_0$ represents an image with the same dimensions as I_0, and $\odot$ represents multiplication. The calculation of the confidence score χ_q^c is shown in Eq. (11).

$$\chi_q^c = F_c(I'_q) - F_c(\tilde{I}_0) \tag{11}$$

In Eq. (11), $F_c(I'_q)$ and $F_c(\tilde{I}_0)$ represent the predicted probability of the image I'_q and $\tilde{I}_0$ in the class c respectively. The calculation of the final saliency map is shown in Eq. (12).

$$L_{Group-CAM}^c = \mathrm{ReLU}\left(\sum_q \chi_q^c M'_q\right) \tag{12}$$

To better preserve the facial features of human faces, research uses facial semantic labels to constrain the facial structure of the generated images. For the acquisition of facial semantic labels, the BiseNet model was selected in this study. The BiseNet model structure mainly includes input images, spatial branch paths, feature fusion modules, output semantic labels and contextual branch paths. The specific structure of the BiseNet model is indicated in Fig. 5 [21].

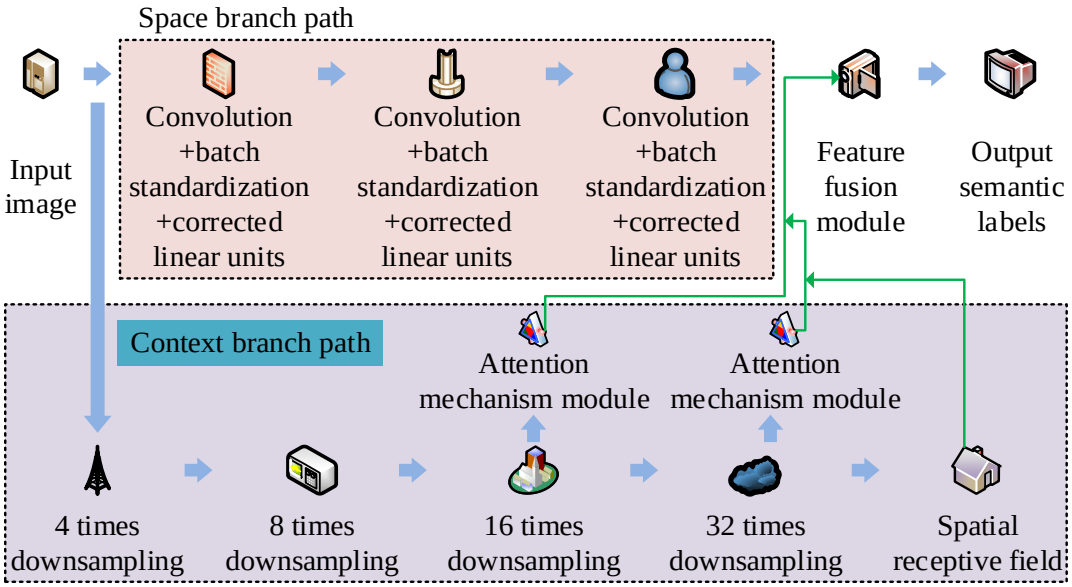


Fig. 5. The specific structure of the BiseNet model.

From Fig. 5, the spatial branch path involves three groups of convolution + batch normalization + modified linear units. The contextual branch path includes 4x, 8x, 16x and 32x downsampling, attention mechanism module and spatial perception wild. In order to allow the generator to retain the characteristics of the comic domain, the study performed strong blur processing on the image, and then calculated the gradient magnitude similarity deviation. The local gradient amplitude is similar as shown in Eq. (13).

$$GMS(\theta)=\frac{2y_u(\theta)y_v(\theta)+c}{y_u^2(\theta)+y_v^2(\theta)+c} \tag{13}$$

Eq. (13), θ is the position of the pixel, which $y_u(\theta)$ denotes the gradient amplitude of the θ pixel in the horizontal direction. $y_v(\theta)$ represents θ the gradient amplitude of the pixel in the vertical direction, which $GMS(\theta)$ is the local gradient field of each small patch . The calculation of gradient amplitude similarity deviation is shown in Eq. (14) [22].

$$GMSD=\sqrt{\frac{1}{T}\sum_{\theta=1}^{T}\left(GMS(\theta)-GMSM\right)^2} \tag{14}$$

Eq. (14), T denotes the total amount of pixels and $GMSM$ denotes the average value of the local gradient field. The loss functions used in the face caricature image generation algorithm include generative adversarial loss, cycle consistency loss, identity loss, class activation mapping loss and weighted sum total loss. The total loss is calculated as denoted in Eq. (15).

$$\min\max \delta_1 L_{lsgan}+\delta_2 L_{cycle}+\delta_3 L_{idenitity}+\delta_4 L_{cam}+\delta_5 L_{gmsd} \tag{15}$$

Eq. (15), δ_1, δ_2, δ_3, δ_4 and δ_5 are all constant weight factors, L_{lsgan}, L_{cycle}, $L_{idenitity}$, L_{cam} and L_{gmsd} respectively represent the generative adversarial loss, cycle consistency loss, identity loss, class activation mapping loss and gradient magnitude similarity bias loss.

IV. Results and Discussion

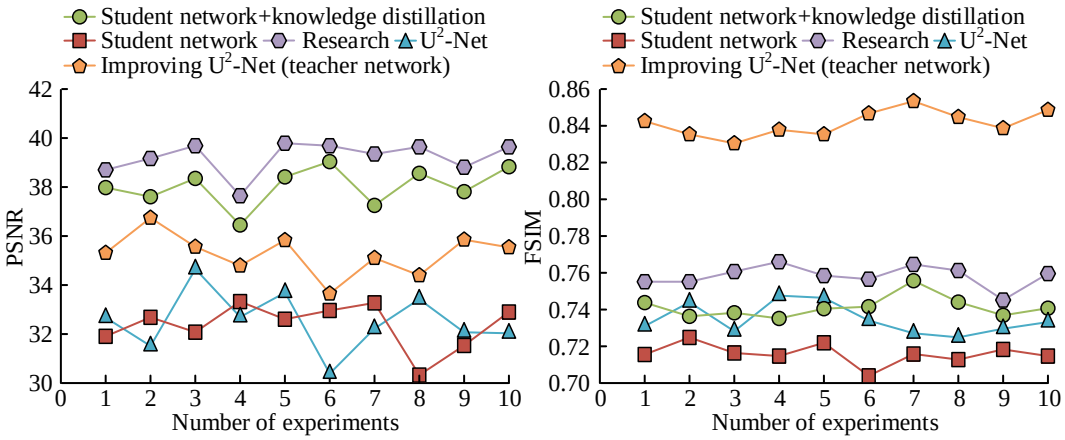
The research verifies the performance of the animation face line drawing generation algorithm and the face caricature image generation algorithm, and explains the data set and experimental environment. The performance verification uses ablation experiments, and uses indicators such as peak signal-to-noise ratio (PSNR), gradient amplitude similarity deviation, and structural similarity to assess the effectiveness of the algorithm.

A. Performance Verification of Animation Face Line Drawing Generation Algorithm

In order to verify the effectiveness of the line drawing generation algorithm, an ablation experiment was conducted. Ablation experiments involve studying the designed model and loss function. Performance evaluation indicators include PSNR, Feature Similarity Index Measure (FSIM), Gradient Magnitude Similarity Deviation (GMSD), structural similarity (Structure Similarity Index Measure, SSIM) and Fréchet Inception Distance (FID). PSNR, FSIM, GMSD, SSIM and FID are all important indicators for measuring image quality. Among them, PSNR is mainly used to compare the differences between the original signal and the processed signal, FSIM is used to quantify the degree of distortion of images in visual perception. GMSD measures the similarity of gradient images and is used to evaluate the clarity of images. SSIM measures the structural similarity between the original image and the processed image, such as brightness, contrast, and structure. FID measures the quality of image generation models. The data set applied in the experiment is the Apdrawing data set, and the algorithm performed a total of 280,000 iterations. In addition, in the line drawing generation algorithm, the values of β_1, β_2, β_3 and β_4 are 10, 1, 1000, and 1/1000, respectively. The operating system applied in the experiment is Windows 11, the processor is Intel Core i9-13900KS, the maximum turbo frequency is

6.00GHz, the basic power consumption of the processor is 150W, the maximum memory is 192GB, the basic frequency and maximum dynamic frequency of the graphics card are 300MHz and 1.65GHz respectively. The comparison of PSNR and FSIM of different models is shown in Fig. 6.

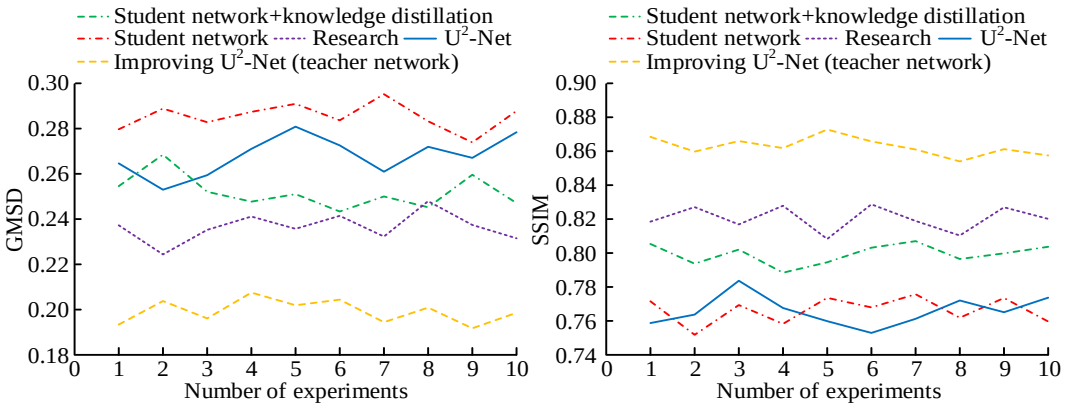
From Fig. 6, the models included in the experiment include the U^2 -Net model, the improved U^2 -Net (teacher network) model, the student network model, the student network + knowledge distillation model and the line drawing generation model designed by the institute. From Fig. 6(a), the max PSNR values of the five models are 34.58, 36.70, 33.55, 38.64 and 39.45 respectively, and the mini values are 31.58, 33.87, 30.33, 36.35 and 37.51 respectively. From Fig. 6(b), the max FSIM values of the five models are 0.7457, 0.8539, 0.7257, 0.7559 and 0.7660 respectively, and the mini values are 0.7224, 0.8305, 0.7066, 0.7351 and 0.7483, respectively. The larger the PSNR and FSIM values are, the better the quality of the images generated by the model is. The PSNR value of the line drawing generation model designed by the institute is significantly greater than the comparison model, which shows that the performance of the model designed by the institute is better. Both the teacher and the student network models after introducing knowledge distillation have improved in PSNR, which can identify the performance of the modules added by the institute. The comparison of GMSD and SSIM of different models is shown in Fig. 7.



(a) Comparison of PSNR values of different models

(b) Comparison of FSIM values of different models

Fig. 6. Comparison of PSNR and FSIM of different models.



(a) Comparison of GMSD values of different models

(b) Comparison of SSIM values of different models

Fig. 7. Comparison of GMSD and SSIM of different models.

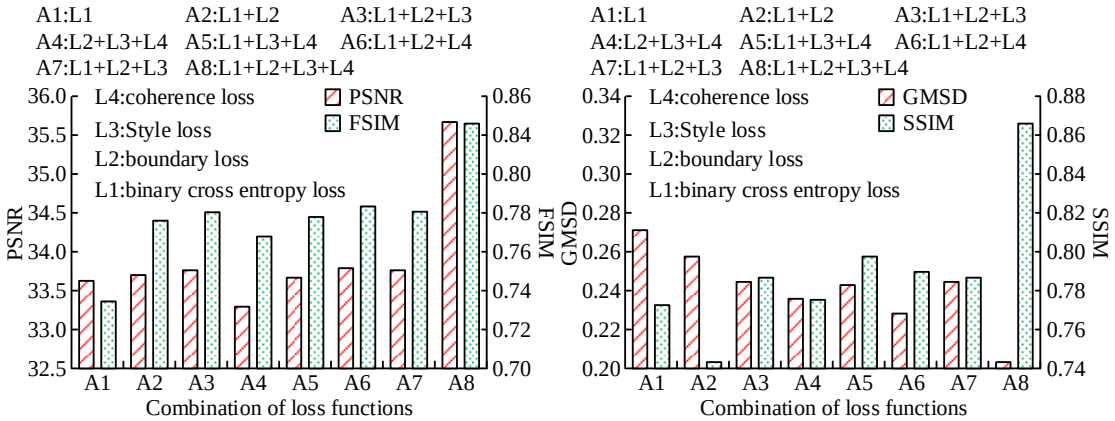
From Fig. 7(a), the max GMSD values of the U^2 -Net model, teacher network model, student network model, student network + knowledge distillation model and research design model are 0.2801, 0.2141, 0.2931, 0.2642 and 0.2432 respectively, the minimum values are 0.2588, 0.1927, 0.2732, 0.2411 and 0.2213, respectively. From Fig. 7(b) that the max SSIM values of the five models are 0.7810, 0.8769, 0.7747, 0.8084 and 0.8285, respectively, and the mini values are 0.7573, 0.8591, 0.7534, 0.7865 and 0.8098 respectively. The larger the GMSD value is, the worse the fidelity of the image is. The larger the SSIM value is, the more similar the image structure is to the real label, and the better the image quality is. The GMSD value and SSIM value of the model designed by the institute have obvious advantages, which also shows that the performance of the model designed by the research is better. The comparison of indicators under different loss function constraints is shown in Fig. 8.

In Fig. 8, the experiment involves a total of 8 combinations of loss functions, which are named A1, A2, A3, A4, A5, A6, A7 and A8 respectively. From Fig. 8(a), the average PSNR of the

eight loss function combinations are 33.58, 33.67, 33.70, 33.29, 33.65, 33.74, 33.70 and 35.70 respectively, and the average FSIM are 0.7346, 0.7752, 0.7880, 0.7679, 0.7783, 0.7814, 0.7880 and 0.8428 respectively. The PSNR value and FSIM value of the loss function combination A8 used in the study are significantly larger than other loss function combinations. From Fig. 8(b), on GMSD, the average values of the eight loss function combinations are 0.2701, 0.2591, 0.2440, 0.2382, 0.2425, 0.2279, 0.2440 and 0.2041 respectively. A8 has the smallest GMSD value. The average SSIM values of the eight loss function combinations are 0.7710, 0.7413, 0.7856, 0.7746, 0.7983, 0.7872, 0.7856 and 0.8669 respectively. A8 has the largest SSIM value. The loss function combination A8 used in the study is beneficial to the image results generated by the final model. In order to better validate the performance of the line drawing generation algorithm, other similar models were selected for comparison. The comparison models include the facial portrait line generation algorithm based on unpaired training data designed by R. Yi et al. [23], the bipartite graph inference generative adversarial network designed by H. Tang et al. [24], and the facial image generation algorithm based on edge optimization and generative adversarial network designed by F. Zhang et al. [25]. The comparison of image generation time and FID using different methods is shown in Table I.

From Table I, it can be seen that in terms of image generation time, the maximum value of the research and design line drawing generation algorithm is 13.88s, and the minimum value is 11.36. The minimum time consumption of facial portrait line generation algorithm based on unpaired training data, bipartite graph inference generative adversarial network, and facial image generation algorithm based on edge optimization and generative adversarial network are 20.62, 23.42, and 16.38, respectively. In addition, the minimum values for the four methods in FID values are 67.52, 115.05, 123.01, and 101.59, respectively. It can be seen that the research and design of line drawing generation algorithms takes less time and the model quality is better. To verify the robustness and generalization ability of the learning network, the image generation effect analysis was conducted. The specific image generation effect is shown in Fig. 9.

From Fig. 9(a) that in the generated image of anime face line drawings, details such as the hair of anime characters are better generated, the lines are smooth and clear, and the features are accurately grasped. The facial features of anime characters are well preserved, such as eyes, noses, etc. In addition, the physical details of anime characters are also well preserved. From Fig. 9(b), when the image generation range is expanded from the face to the whole body, the generated line drawing image effect is also very good, and the hair, charm, body structure and other characteristics of the anime characters are well preserved. The algorithm designed by the research has good generalization ability and robustness.



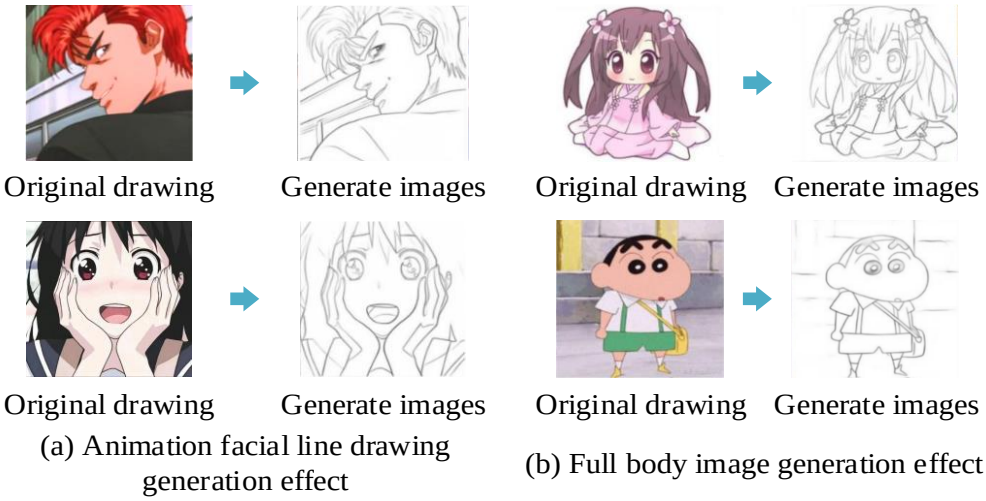
(a) Comparison of PSNR and FSIM values for different loss function combinations

(b) Comparison of GMSD and SSIM values for different loss function combinations

Fig. 8. Comparison of indicators under different loss function constraints.

TABLE I. COMPARISON OF IMAGE GENERATION TIME AND FID USING DIFFERENT METHODS

Model	Time consumption/s					FID				
	Number of experiments					*Number of experiments*				
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
R. Yi et al. [23]	21.36	20.62	21.37	22.09	21.75	115.05	117.84	121.10	117.78	116.59
H. Tang et al. [24]	25.42	24.17	24.04	25.83	23.42	123.01	125.24	127.70	130.51	129.32
F. Zhang et al. [25]	17.87	16.38	17.16	18.33	19.01	104.23	103.83	109.78	101.59	106.60
Manuscript	12.97	11.71	13.88	12.55	11.36	67.52	69.04	68.54	70.13	68.66



(a) Animation facial line drawing generation effect

(b) Full body image generation effect

Fig. 9. Specific image generation effects.

B. *Performance Verification of Face Caricature Image Generation Algorithm*

To assess the effectiveness of the face caricature image generation algorithm, an ablation experiment was conducted. Ablation experiments are mainly carried out from the overall model. The overall model compares U-GAT-IT (baseline), U-GAT-IT+semantic constraints and the comic image generation model designed by the research institute [26]. The indicators used in the experiment include FID, Mean Squared Error (MSE), PSNR and SSIM. Among them, MSE is an indicator used to measure the difference between a model's predicted values and actual observed values, and is commonly used to evaluate the degree of fit of a model on a given data. The data sets used in the experiment include the Flickr-Faces-High-Quality (FFHQ) data set and the Avatar data set. The operating system and processor used in the experiment are the same as those in Section III(A) and will not be repeated here. The facial comic image generation algorithm uses an Adam optimizer with a learning rate of le-4 and a training batch size of 1. In addition, the values of the algorithm on δ_1, δ_2, δ_3, δ_4, and δ_5 are 1, 10, 10, 1000, and 10 respectively. The comparison of FID values and MSE values of different models is shown in Table II.

From Table II, it can be seen that the maximum FID values of the U-GAT-IT model, U-GAT-IT + semantic constraints and the comic image generation model designed by the institute are 144.68, 103.49 and 81.60 respectively, and the minimum values are 139.54, 139.54 and 81.60 respectively. The FID index can express the similarity of feature distributions of two sets of images, and the smaller the FID value, the more similar the feature distributions are. In addition, the max MSE values of the three models are 3.27, 2.98, and 1.42 respectively, and the mini values are 3.04, 2.65, and 1.21 respectively. The MSE metric can also evaluate the quality of images generated by the model. The FID value and MSE value of the model designed by the institute are significantly lower than the baseline model, and the FID value and MSE value of the U-GAT-IT + semantic constraint model are also significantly lower than the baseline model. This shows that the comic image generation model designed by the research has better performance, and also proves the effectiveness of the semantic constraints and group activation mapping modules. The comparison of PSNR values and SSIM values of different models is indicated in Table III.

From Table III, the max PSNR values of the U-GAT-IT model, U-GAT-IT + semantic constraints and the research design model are 32.65, 36.97 and 39.65 respectively, and the mini values are 31.87, 35.36 and 38.44 respectively. In terms of SSIM values, the max values of the three models are 0.7357, 0.7743 and 0.8284 respectively, and the mini values are 0.7123, 0.7615 and 0.8117, respectively. The PSNR value and SSIM value of the model designed by the institute are significantly greater than those of the baseline model and the U-GAT-IT + semantic constraint model, which shows that the performance of the model designed by the institute is better and the quality of the images it generates is better. To better verify the performance of the model designed in the study, the study selected other advanced unsupervised models for comparison. Comparative models include Cycle-consistent Generative Adversarial Network (CycleGAN), Adaptive Convolutions (AdaConv) and No Independent Component Encoding Generative Adversarial Network (NICEGAN). The comparison of FID and time consumption of different models is shown in Fig. 10.

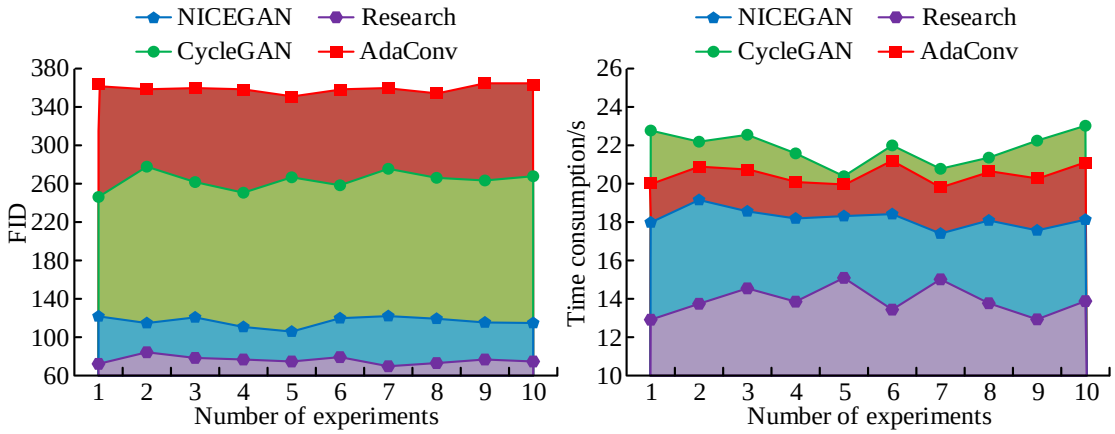
From Fig. 10(a), in terms of FID values, the maximum values of CycleGAN, AdaConv, NICEGAN and the research design model are 263.57, 365.96, 119.47 and 81.60 respectively, and the minimum values are 251.75, 352.64, 102.31 and 71.32 respectively. From Fig. 10(b), in terms of time consumption, the maximum values of the four models are 22.54s, 21.31s, 19.32s and 15.21s respectively, and the minimum values are 20.46s, 19.89s, 17.65s and 13.24s respectively. Whether it is in terms of FID value or model time consumption, the performance of the research design model has more advantages.

TABLE II. COMPARISON OF FID AND MSE VALUES FOR DIFFERENT MODELS

Model	FID					MSE				
	Number of experiments					*Number of experiments*				
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
U-GAT-IT	139.54	140.87	144.68	143.92	142.85	3.27	3.11	3.21	3.04	3.18
U-GAT-IT +semantic constraints	97.45	103.49	102.71	99.21	95.86	2.65	2.78	2.98	2.82	2.73
Research	71.32	75.64	81.60	73.17	77.48	1.42	1.37	1.29	1.32	1.21

TABLE III. COMPARISON OF PSNR AND SSIM VALUES FOR DIFFERENT MODELS

Model	PSNR					SSIM				
	Number of experiments					*Number of experiments*				
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
U-GAT-IT	32.57	32.24	31.98	32.65	31.87	0.7123	0.7344	0.7357	0.7224	0.7234
U-GAT-IT +semantic constraints	35.82	36.43	35.64	36.97	35.36	0.7647	0.7743	0.7684	0.7718	0.7615
Research	38.75	39.46	38.44	39.65	39.13	0.8257	0.8117	0.8226	0.8273	0.8284



(a) Comparison of FID values for different models

(b) Comparison of time consumption of different models

Fig. 10. Comparison of FID and time consumption for different models.

C. Discussion

Aiming at the improvement of image quality generated by anime drawing robots, this study designs facial line drawing generation algorithms and comic image generation algorithms from the perspectives of anime lines and comic images. The results show that the maximum PSNR of the knowledge distillation based generation algorithm is 39.45, and the minimum value is 37.51, which is significantly better than the comparison model. Researchers such as M. Yuan have designed a cross task knowledge distillation method and a multi-stage knowledge distillation paradigm to address the issue of text synthesized images, achieving improvements in visual quality and semantic consistency of synthesized images [27]. The generation algorithm based on knowledge distillation is similar to the research results of M. Yuan et al. The maximum and minimum FID values of the comic image generation model based on semantic constraints are 81.60 and 71.32, respectively, with a maximum time consumption of 15.21 seconds. The performance is superior to the comparison model. In order to solve the problem of low image generation quality under limited data, Y. Gou et al. designed a cross domain semantic relationship loss to improve the performance of image generation models under limited data. The comic image generation model based on semantic constraints is similar to the research results of Y. Gou et al. [28].

V. CONCLUSION

To raise the quality of images generated by animation drawing robots, an animation face line drawing generation algorithm based on knowledge distillation was designed, and an unsupervised face comic image generation algorithm based on semantic constraints was also designed. The results show that the maximum PSNR values of the U^2 -Net model, teacher network model, student network model, student network + knowledge distillation model and line drawing generation model

are 34.58, 36.70, 33.55, 38.64 and 39.45 respectively, and the minimum values are 31.58, 33.87, 30.33, 36.35 and 37.51. The performance of the line drawing generation model designed by the institute is better, and the modules added by the institute are effective. The average values of the loss functions PSNR, FSIM, GMSD and SSIM of the line drawing generation model are 35.70, 0.8428, 0.2041 and 0.8669 respectively. Investigate the combinations of loss functions used that are beneficial to the image results generated by the final model. The maximum FID values of the U-GAT-IT model, U-GAT-IT + semantic constraints and comic image generation model are 144.68, 103.49 and 81.60 respectively, and the minimum values are 139.54, 95.86 and 71.32 respectively. The maximum and minimum time consumption of the comic image generation model are 15.21s and 13.24s respectively. The performance of the comic image generation model is better, and the semantic constraints and group activation mapping modules used in the study are effective. The performance of the comic image generation of research model can be further improved on some images. Future research can introduce the Spade module to maintain the structure and improve the quality of image generation on facial features. In addition, future research can also extend knowledge distillation to multi-task models to improve the performance of learning network models.

REFERENCES

[1] Y. Song, J. Wu, Z. Liu, B. Zhang, and T. Huang, "Similitude analysis method of the dynamics of a hybrid spray-painting robot considering electro-mechanical coupling effect," IEEE-ASME. T. Mech, vol. 26, no. 6, pp. 2986-2997, January 2021.

[2] A. Muneer, and Z. Dairabayev, "Design and implementation of automatic painting mobile robot," IAES Int. J. Robot. Autom., vol. 10, no. 1, pp. 68-74, March 2021.

[3] N. Yu, L. Nan, and T. Ku, "Robot hand-eye cooperation based on improved inverse reinforcement learning," Ind. Robot, vol. 49, no. 5, pp. 877-884, June 2022.

[4] Y. Liu, A. Ojha, S. Shayesteh, H. Jebelli, and S. H. Lee, "Human-centric robotic manipulation in construction: generative adversarial networks based physiological computing mechanism to enable robots to perceive workers' cognitive load," Can. J. Civil. Eng, vol. 50, no. 3, pp. 224-238, February 2023.

[5] P. P. Groumpos, "A critical historic overview of artificial intelligence: issues, challenges, opportunities, and threats", Artif. Intell. Appl, vol. 1, no. 4, pp. 197-213, June 2023.

[6] D. K. Ko, D. H. Lee, and S. C. Lim, "Continuous image generation from low-update-rate images and physical sensors through a conditional gan for robot teleoperation," IEEE. T. Ind. Inform, vol. 17, no. 3, pp. 1978-1986, May 2021.

[7] R. Liu, W. Wan, K. Koyama, and K. Harada, "Robust robotic 3-D drawing using closed-loop planning and online picked pens", IEEE. T. Robot, vol. 38, no. 3, pp. 1773-1792, June 2021.

[8] Z. Khanam, B. Aslam, S. Saha, X. Zhai, and K. Mcdonald-Maier, "Gamma-induced image degradation analysis of robot vision sensor for autonomous inspection of nuclear sites," IEEE. Sens. J, vol. 22, no. 18, pp. 17378-17390, January 2021.

[9] T. Wang, W. Q. Toh, H. Zhang, X. Sui, and W. Jing, "RoboCoDraw: Robotic avatar drawing with gan-based style transfer and time-efficient path optimization," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 6, pp. 10402-10409, April 2020.

[10] P. L. Wu, Y. C. Hung, and J. S. Shaw, "Artistic robotic pencil sketching using closed-loop force control," Proceedings of the Institution of Mechanical Engineers, Part C. Journal of Mechanical Engineering Science, vol. 236, no. 17, pp. 9753-9762, May 2022.

[11] P. Jens, and R. Dagmar, "Robotic drawing communication protocol: a framework for building a semantic drawn language for robotic fabrication," Construc. Robotics, vol. 6, no. 3, pp.239-249, January 2022.

[12] L. Scalera, E. Maset, S. Seriani, A. Gasparetto, and P. Gallina. "Performance evaluation of a robotic architecture for drawing with eyes," Int. J. Mech. Con, vol. 22, no. 2, pp. 53-60, April 2021.

[13] E. W. Herrmann, and A. Bigham, "Drawing fields: prototyping public space with semi-autonomous robots," Int. J. Archit. Comput, vol. 19, no. 4, pp. 612-617, August 2021.

[14] H. Pranoto, Y. Heryadi, H. L. H. S. Warnars, and W. Budiharto, "Enhanced IPCGAN-Alexnet model for new face image generating on age target," J. King. Saud. Univ-Com, vol. 34, no. 9, pp. 7236-7246, September 2022.

[15] X. Tu, Y. Zou, J. Zhao, W. Ai, and J. Feng, "Image-to-Video Generation via 3D Facial Dynamics," IEEE. T. Circ. Syst. Vid, vol. 32, no. 4, pp. 1805-1819, April 2022.

[16] J. Yang, Y. Wang, H. Zao, and G. Gui, "MobileNet and knowledge distillation-based automatic scenario recognition method in vehicle-to-vehicle systems," IEEE. T. Veh. Technol, vol. 71, no. 10, pp. 11006-11016, October 2022.

[17] H. Salman, A. H. Taherinia, and D. Zabihzadeh, "Fast and accurate image retrieval using knowledge distillation from multiple deep pre-trained networks," Multimed. Tools. Appl, vol. 82, no. 22, pp. 33937-33959, March 2023.

[18] Z. Feng, X. Cui, H. Ji, M. Zhu, and L. Stankovic, "VS-CAM: vertex semantic class activation mapping to interpret vision graph neural network," Neurocomputing, vol. 533, no. 7, pp. 104-115, September 2023.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," INT. J. Comput. Vision, vol. 128, no. 2, pp. 336-359, February 2020.

[20] J. S. Yun, and S. B. Yoo, "Kernel-attentive weight modulation memory network for optical blur kernel-aware image super-resolution," Opt. Lett, vol. 48, no. 10, pp. 2740-2743, May 2023.

[21] T. H. Tsai, and Y. W. Tseng, "BiSeNet V3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation," Neurocomputing, vol. 532, no. 1, pp. 33-42, February 2023.

[22] A. Paul, "Adaptive tri-plateau limit tri-histogram equalization algorithm for digital image enhancement," Visual. Comput, vol. 39, no. 1, pp. 297-318, November 2023.

[23] R. Yi, Y. J. Liu, Y. K. Lai, and P. L. Rosin, "Quality Metric Guided Portrait Line Drawing Generation from Unpaired Training Data," IEEE. T. Pattern Anal., vol. 45, no. 1, pp. 905-918, January 2023.

[24] H. Tang, L. Shao, P. H. S. Torr, and N. Sebe, "Bipartite Graph Reasoning GANs for Person Pose and Facial Image Synthesis," Int. J. Comput. Vision, vol. 131, no. 3, pp. 644-658, December 2023.

[25] F. Zhang, H. Zhao, W. Ying, Q. Liu, and B. Fu, "Human Face Sketch to RGB Image with Edge Optimization and Generative Adversarial Networks," Intell. Autom. Soft. Co, vol. 26, no. 6, pp. 1391-1401, January 2020.

[26] N. Yang, B. Xia, Z. Han, and T. Wang, "A domain-guided model for facial cartoonlization," IEEE/CAA. J. Autom. Sinica, vol. 9, no. 10, 1886-1888, August 2022.

[27] M. Yuan, and Y. Peng, "CKD: Cross-Task Knowledge Distillation for Text-to-Image Synthesis," IEEE. T. Multimedia, vol. 22, no. 8, pp. 1955-1968, August 2020.

[28] Y. Gou, M. Li, Y. Lv, Y. Zhang, Y. Xing, and Y. He, "Rethinking cross-domain semantic relation for few-shot image generation," Appl. Intell, vol. 53, no. 19, pp. 22391-22404, June 2023.