ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

High-Resolution Locally Controllable Portrait Style Transfer

Guoquan Jiang[1,a], Huan Xu[2,b,*], Zhanqiang Huo[1,c]

[1]School of Software, Henan Polytechnic University, Jiaozuo, 454000, China
[2]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454000, China
[a]jiangguoquan@hpu.edu.cn, [b]212209020020@home.hpu.edu.cn, [c]hzq@hpu.edu.cn
**Corresponding author: 212209020020@home.hpu.edu.cn*

Abstract: *Most of the current portrait style transfer algorithms focus on the overall portrait style transfer for low-resolution images. However, in real life, users need to have fine control over specific regions of images with different resolutions, and the stylized images generated by existing methods have problems such as structure loss, local contour deformation, and color rendering errors. Therefore, this paper proposes a high-resolution locally controllable portrait style transfer model. By introducing a novel U-block, the method is not only suitable for local portrait style transfer but also can effectively handle the overall portrait style transfer task. By adopting two U-shaped encoders with different structures, this method constructs a unique generator structure, so that the structural features of the content domain and the style domain can be learned more fully, and the problem of structure loss in the stylized image is reduced. In addition, we propose a local portrait style transfer module, which allows users to perform accurate local style transfer based on the segmentation masks of different regions. To further improve the effect of local feature fusion and reduce the distortion of local contour, a Local feature fusion module (LFFM) is designed, which abandons the traditional feature splicing method and improves the quality of local stylized images by using a style attention mechanism. Finally, to reduce artifacts and color rendering errors, a local portrait style loss is introduced as a constraint to ensure that the style transfer region accurately learns the target style, while keeping the original structural features of other regions unchanged by histogram matching. The results of comparative and ablation experiments on four different style datasets show that the proposed method achieves excellent performance in global and local portrait style transfer, which verifies the effectiveness of the proposed method in portrait style transfer.*

Keywords: *portrait style transfer; U-shaped encoder; Local feature fusion module; Local portrait style loss*

1. Introduction

Style transfer techniques have been extensively studied in non-photorealistic rendering (NPR)[1] and deep learning research[2], and in the NPR field, many approaches[3-4] have been investigated for automatic portrait generation. Based on StyleGAN[5-6], the framework of the generated high-resolution art portrait has made very good progress. Kong et al.[7] proposed an asymmetric two-stream generative adversarial network ADS-GAN to solve the problems of facial deformation and contour loss caused by cartoons and other style transfer techniques applied to portrait images. However, this method can only transfer the overall style of the style dataset, and cannot realize the style transfer of individual styles. Song et al.[8] proposed AgileGAN to generate a high-quality style portrait framework by inversion consensus learning and introduced a new variational autoencoder that can capture more different levels of detail. Although this method can be used for portrait style transfer of high-resolution images, the generated stylized images modify too many detailed features of the face. Yang et al.[9] introduced an extrinsic path with StyleGAN[5-6] as the framework design, which can realize the overall portrait style transfer by using unpaired data sets. Wang et al.[10] proposed a facial pose awareness and style transfer network Face-PAST, which preserves facial structure and details by using a pre-trained style generation network in external style transfer by improving the method of Yang et al. [9] Qi et al.[11] introduced DEADiff to solve the problem that the text controllability of the text-to-image model was seriously damaged when the existing encoder method transmitted the ring style. Although this method no longer needed the artistic style data set, it also made the portrait style transfer only applicable to specific fields, and the overall stylized image could be obtained only by accurate description. It cannot achieve controllable local portrait style transfer for specific regions.

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

In summary, although the existing portrait style transfer model can generate portrait style images with good effects, it can only learn the overall target style and cannot perform local portrait style transfer of accurate parts. Especially for high-resolution images, it usually requires large memory requirements and powerful computing power support, which becomes a big problem for local portrait style transfer. Moreover, the stylized images generated by existing algorithms suffer from facial distortion, structure loss, and artifacts. To solve the above problems, this paper proposes a high-resolution locally controllable portrait style transfer algorithm. In the portrait style transfer task, the proposed model can achieve good results both in the integrity of the portrait contour of the reconstructed content image and the learning of highly abstract style features. Problems such as loss of details, distortion of facial structures, and artifacts in stylized images are greatly reduced. Considering the huge difference between the content domain and the style domain, which is different from the traditional style transfer model, we design the input structure of the generator as two feature encoders with different structures, so that the network can better extract the content features of the content image and the style features of the style image. In addition, a local portrait style transfer module is designed to enable the network architecture to realize local portrait style transfer of accurate regions. The local portrait style loss and cycle consistency loss are designed to constrain the facial contour of the stylized image, which reduces the loss of the structure of the content image and reduces the generation of artifacts. In summary, our main contributions are as follows:

1) This paper proposes a high-resolution locally controllable portrait style transfer algorithm, which can not only be applied to high-resolution images but also improve the quality of portrait style transfer without increasing the memory and computation of the model. Aiming at the huge difference between the content domain and the style domain, two U-shaped feature encoders with different structures are used to transfer the portrait style of the high-resolution unpaired artistic style images, to improve the quality of the overall portrait style transfer.

2) To increase the feature extraction ability of the input image, the input of the generator is designed into two U-shaped encoders with different structures. In addition, to realize local portrait style transfer, a local portrait style transfer module is designed. The network structure can accurately divide different regions of the portrait image. According to the segmentation masks of different regions selected by the user, the regions that need local portrait style transfer are fused by the Local feature fusion module (LFFM). Histogram Matching is performed on the parts without local portrait segmentation transfer to preserve the structure of the original content image. It not only realizes portrait style transfer for specific regions but also retains the structure of other original content images, reducing the problem of facial structure distortion and artifacts in stylized images.

3) In addition, in order to increase the accuracy of local portrait style transfer, we also design a local portrait style loss L_{local_style}. In order to prevent the deformation of stylized facial structure, we use the cycle consistency loss L_{cyc} to constrain the image contour, which further improves the quality of portrait style transfer.

4) Through a series of experiments on four different styles of data sets, various comparative experiments prove that the method in this paper is superior to the current most advanced methods.

2. Related work

Isola et al.[12] first proposed a supervised image translation framework, which can map the source domain to the target domain and easily achieve style transfer. However, this framework requires paired data for training, which is a very difficult thing for portrait style transfer style. Zhu et al.[13] proposed cycle consistency loss for unsupervised image translation, which makes it possible to perform style transfer tasks without using paired datasets. Chen et al.[14] improved Isola et al.[12] method using unpaired data training by combining neural style transfer and generative adversarial networks, which can quickly convert real images into animation images. Gatys et al.[15] utilized a pre-trained network architecture VGG-19 as a feature extractor to drive texture synthesis and style transfer, exhibiting rich feature representations and thus gaining extensive attention in the academic community. Xing et al.[16] proposed a portion-aware artistic style transfer method, which can separate the portrait from the background by generating a background mask with the help of semantic segmentation content image. However, this method can only transfer the overall target style. Qin et al.[17] proposed a U^2 architecture, which can make the network deeper and achieve higher resolution. While ensuring high resolution, it does not increase memory and computational cost. The nested U-shaped structure can extract multi-scale features at the bottom without reducing the resolution of the feature map. Each stage is filled by a U-shaped residual block, which can segment the target area more accurately. Therefore, this structure is

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

very promising in the direction of high-resolution local portrait style transfer.

3. Method

3.1 Overview

The high-resolution locally controllable portrait style transfer algorithm proposed in this paper is based on the GAN framework. The biggest difference from the traditional GAN is that the network architecture of this paper is a dual-input/dual-output structure. Considering the huge difference between the source and target domains, two U-shaped encoders with different structures are included in the generator in this paper, which are used to extract the features of different input images respectively. The images used in this paper can be unpaired images of any resolution size, where the content and style domains are denoted as $T \subset R^{H\times W\times 3}$ and $S \subset R^{H\times W\times 3}$, respectively. Our model contains two generators with different structures and two discriminators with the same structure, namely $G_{S\to T}$, $G_{T\to S}$, D_S and D_T. $S \to T$ represents the mapping process from the content image to the style image domain, and $T \to S$ represents the mapping process from the style image to the content image domain. D_S is used to identify whether the generated image belongs to the S domain, and D_T is used to identify whether the generated image belongs to the T domain. Compared with traditional methods, our model adds a reverse mapping process to realize the cyclic mapping of the model. The overall goal is to transfer the abstract elements of the abstract artistic style to the characters in the target portrait image, but the biggest difference with the traditional portrait style transfer is that Our method can realize not only the overall artistic portrait style transfer but also the local portrait style transfer. The overall model structure is shown in Fig 1.

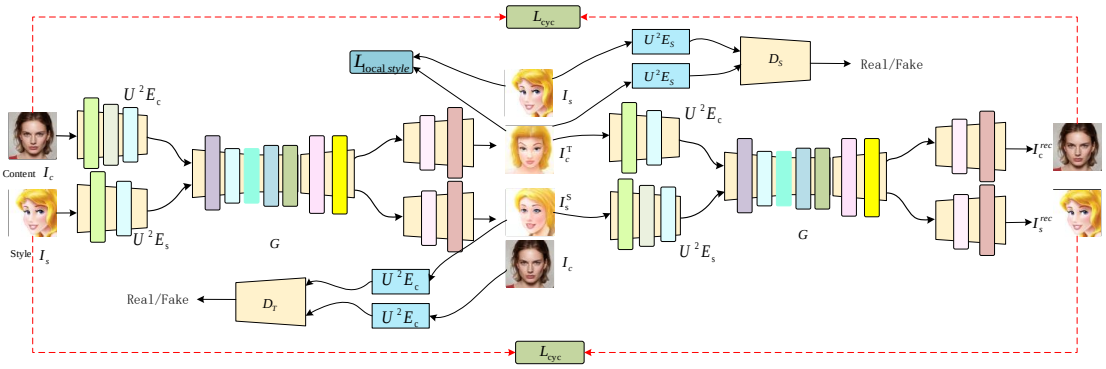

Fig 1. High-resolution locally controllable portrait style transfer network framework.

Our network architecture works by learning a bidirectional mapping between two domains, namely G: $S \times T \to T \times S$. Specifically, the model receives a portrait image $I_c \in S$ and a reference artistic style image $I_s \in T$ as input to the generator G. After training and learning, the model can generate an artistic style image I_c^s that only transfers local regions (e.g., mouth, nose, eyes, etc.), and I_c^s has the structural characteristics of $I_c \in S$. At the same time, the invariance of the region without portrait style transfer is maintained, and the overall artistic style characteristics of $I_s \in T$ are preserved. The specific generation process of the model is as follows:

$$(I_c, I_s) \to G_{S\to T}(I_c, I_s) \to (I_c^T, I_s^S) \tag{1}$$

$$(I_c^T, I_s^S) \to G_{T\to S}(I_c^T, I_s^S) \to (I_c^{rec}, I_s^{rec}) \tag{2}$$

$$(I_c^{rec}, I_s^{rec}) \to (I_c, I_s) \tag{3}$$

where I_c and I_s denote the content image and artistic portrait style image respectively, I_c^T and I_s^S denote the artistic portrait style image corresponding to the content image and the artistic portrait style image corresponding to the content image respectively, and I_c^{rec} and I_s^{rec} denote the reconstructed content image and artistic portrait style image respectively.

To make the content encoder and style encoder more fully extract the structural information from the image from multiple scales, the generator frame in the traditional method is improved. By introducing the U^2— Net[17] structure, the input of the generator is designed as a U-shaped encoder (U^2E_c and U^2E_s). It can capture more useful feature information from input content images and artistic portrait photos at multiple different scales. In addition, to facilitate the subsequent local portrait style transfer of

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

sub-regions, the region in the content image is segmented into different regions, each region has a mask, and then the semantic information of the corresponding region is extracted in turn. Finally, the regional mask is specified, and the portrait style transfer is carried out by referring to the style image to complete the process of local portrait style transfer.

3.2 Generator structure

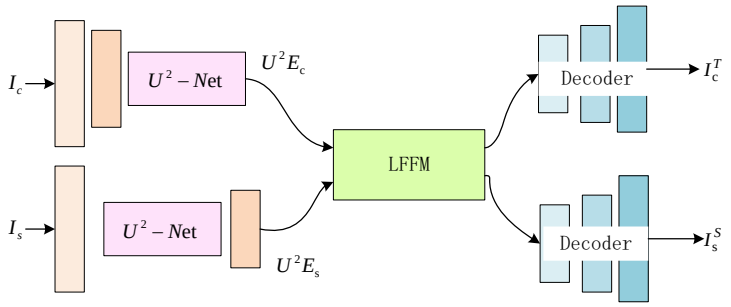

Fig 2. Generator structure.

Our model designs the structure of the generator as a two-input/two-output structure according to the large difference between the source domain and the target domain. Specifically, the generator contains two U-shaped encoders, which are the content encoder U^2E_c and the style encoder U^2E_s, a local feature fusion module (LFFM) and two decoders with the same structure, as shown in Fig. 2. Inspired by the structure of SA-Net[18], the Local Feature fusion module (LFFM) in this paper adopts the style attention mechanism, but the biggest difference is that the number of input features is different, including local content features and local style features respectively, which can increase the effect of local portrait style transfer and retain more content structure. The architecture of both decoders is the same and consists of a series of convolutional layers combined.

The generator $G_{S\to T}$ receives content image $I_c \in S$ and art portrait style image $I_s \in T$ as inputs at the same time. After model training, the stylized image I_c^T contains both style features of the style image and content features of the content image. The process can be expressed as follows:

$$(I_c^S, I_s^T) \to G_{S\to T}(I_c, I_s) \tag{4}$$

Because the style of the four target artistic portrait style datasets in this paper is highly abstract, the convolution kernel size of the conventionally designed CNN[15] or RseNet[19] is usually 1×1 or 3×3, which is easy to ignore important structural information during feature extraction, resulting in the loss of important features in the stylized image. More importantly, when used for local portrait style transfer, a small receptive field easily leads to the loss of local detail features, which in turn leads to local distortions in the stylized image of local portrait style transfer. Inspired by U^2— Net[17], this chapter sets the encoder of the generator to two different structures, namely the content encoder U^2E_c and the style encoder U^2E_s. The unique U-shaped design significantly enhances the network's ability to extract important features of the image, which not only ensures the quality of local portrait style transfer but also improves the performance of the network. It also improves the stylization effect of the overall portrait style transfer.

3.3 U-shaped feature encoder

In the task of portrait style transfer, both global feature information and local feature information are crucial, especially for local portrait style transfer. If the local feature information cannot be accurately obtained, there will be inaccurate segmentation regions, which will lead to serious facial artifacts and distortions in the local portrait stylized image. Therefore, to more fully extract important feature information from the input content image and style image, this chapter introduces the U^2-Net structure[17] to design the U-shaped feature encoder structure in the model. Different from the receptive fields of 1×1 or 3×3 size in traditional convolutional neural networks, U^2-Net[17] extracts the key features of input content images and style images from multi-scale through receptive fields of different sizes. The pooling operation can extend the depth of the model so that the model does not increase additional computational cost when processing high-resolution images. The specific structure of the U^2-Net model is shown in Fig 3.

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

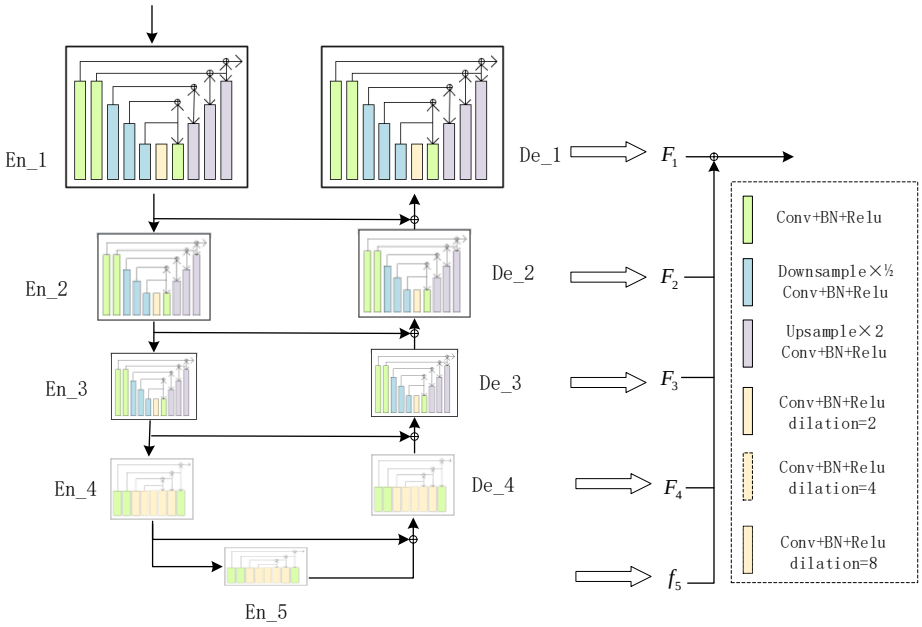

Fig 3. Specific structure of $U^2 - Net$ model in this paper.

In the style encoder U^2E_s and the content encoder U^2E_c, the input content image and style image are first subjected to a preliminary convolution operation to extract the underlying important features of the image, which can be expressed by the following formula:

$$f_0 = conv(1) \tag{5}$$

After the feature map f_0 is passed into the encoder module, it needs to continue to extract the features in the high-level layer. Different from the 11 stages contained in the original $U^2 - Net$[17] structure, the $U^2 - Net$ structure we adopted consists of a total of 9 stages: a 5-stage encoder (En_1, En_2, En_3, En_4, En_4), and a 4-stage decoder (De_1, De_3, De_3, De_4). The specific process can be described as follows:

$$f_i = En\_i(f_{i-1})\ i = 1,2,3,4,5 \tag{6}$$

$$F_j = \begin{cases} De\_j(f_i, F_{j+1})\ j = 1,2,3 \\ De\_j(f_i, f_{j+1})\ j = 4 \end{cases} \tag{7}$$

$$F = concat(F_1, F_2, F_3, F_4, f_5) \tag{8}$$

Here, i and j denote the number of encoder and decoder stages in $U^2 - Net$, respectively. Among them, the encoder stages En_1, En_2, and En_3 are composed of residual U-shaped blocks of different heights, and the decoder stages De_1, De_2, and De_3 are similar to the encoder stage, as shown in Fig. 3. Since the feature maps in En_4, En_5, and De_4 have relatively low resolution, to avoid the loss of important features, we use dilated convolution to replace the pooling operation and sampling operation in the original residual U-block, that is, En_4, The resolution of the feature maps produced by all intermediate processes in En_5 and De_4 is the same as the discrimination rate of the input features they initially received. Therefore, the multi-scale features within a stage and the aggregation of multi-scale features between stages can be more effectively extracted through this structure, which is convenient for subsequent portrait style transfer operations.

3.4 Local portrait style transfer network

A very key place in the local portrait style transfer is the local feature fusion module (LFFM). Before and after the portrait style transfer, it is necessary to ensure that the part of style transfer is distinguished from the part of the portrait without style transfer, and only the specific region is transferred, while the rest of the structure is kept unchanged. Histogram Matching is used on the pixel-level feature map to ensure that the output stylized image is consistent with the content image and the reference artistic style image features.

Firstly, the content image I_c and the style image I_s are fed to the U-shaped feature encoder for

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

region segmentation, and the masks of different regions in the content image and the artistic style image are obtained. Then the eyebrows, eyes, nose, mouth, and hair are picked out. The fully segmented face mask, including the neck, is then binary masked to form new masks: $M_{eyebrow}$, M_{eye}, M_{nose}, M_{lip}, M_{face}, and M_{hair}. The required eye mask M_{eye} was extracted, and the region of the image to be transferred for portrait style was extracted by multiplying the M_{eye} segmented by the content image and the style image. The local feature fusion module (LFFM) was used for local style transfer. The regions without portrait style transfer are processed by color-based Histogram Matching, and the local stylized image $H(x,y)$ is obtained, which is consistent with the structure distribution of the original content image and has the style image style. To prevent color loss and artifacts in the process of local portrait style transfer, the local style loss function L_{local_style} is used to impose constraints between the output local stylized image $H(x,y)$, the content image I_c and the artistic style image I_s.

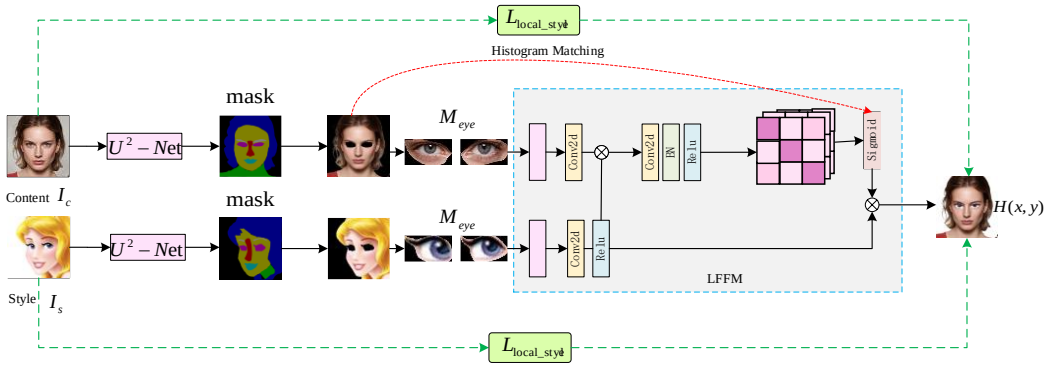

Fig 4. Local portrait style transfer network.

3.5 Objective loss function

Our main goal is how to achieve local portrait style transfer while improving the overall portrait style transfer and solving the problems of structure loss, facial distortion, and artifacts in the stultification results of current portrait style transfer algorithms. Therefore, we design the loss function from the perspective of portrait structure preservation, local style feature learning, and global structure feature preservation. Firstly, from the aspect of target style learning of the network model, we design a local portrait style loss L_{local_style}, so that the content image can learn the style features of the target style image while retaining the content image content features during the local portrait style transfer. Secondly, from the perspective of portrait results, to maintain the stability of local portrait style transfer results, the cyclic consistency loss is used to constrain the consistency between the reconstructed image and the original input image, so it also maintains the effect stability of local portrait style transfer to a certain extent. Through the cooperation of various losses, the model in this chapter not only performs well on local portrait style transfer but also significantly improves the overall portrait style transfer effect on four artistic portrait styles.

(1) Local portrait style loss

For a given input content image I_c and style image I_s, to make the output local stylized image $H(x,y)$ not only learn the style features of the target style image but also retain the content features of the content image, the input content image I_c and style image I_s are input into the model to extract feature maps. The similarity between the structural features of the local stylized image $H(x,y)$, the content image I_c, and the style image I_s can be obtained by calculating the Gram matrix. Specifically, the local portrait style loss function of layer i is expressed as follows:

$$L_{local\_style} = \frac{1}{C_l \times H_l \times W_l} \sum_{ijk} [G_l(H(x,y)) - G_l(I_c, I_s)]^2_{ijk} \quad (9)$$

$$G_l(\cdot) = F_{l,m}(\cdot)F_{l,m}(\cdot)^T, F_{l,m}(H(x,y)) = F_l(H(x,y))M_{l,m}(I_c,I_s), F_{l,m}(I_c,I_s) = F_l(I_c,I_s)M_{l,m}(I_c,I_s) \quad (10)$$

Here, the Gram matrix of the lTH layer feature map is denoted by $G_l(\cdot)$, and m represents a certain region of the portrait semantic segmentation M.

(2) Adversarial loss

The adversarial loss consists of two parts, mainly composed of D_S and D_S:

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

$$L_{adv} = L_{D_S} + L_{D_T} \tag{11}$$

First, the goal of D_S is to distinguish the generated I_c^T from the real portrait photo I_c. The formula is expressed as follows:

$$L_{D_s} = E_{I_c}[logD_s(U^2E_c(I_c)) + E_{I_c,I_s}[log(1 - D_s(U^2E_c(I_c^T)))] \tag{12}$$

Similarly, the goal of D_T is to distinguish the generated image I_s^S from the reference artistic portrait I_s, and the formula is expressed as:

$$L_{D_T} = E_{I_s}[logD_T(U^2E_s(I_s)) + E_{I_c,I_s}[log(1 - D_T(U^2E_s(I_s^S)))] \tag{13}$$

(3) Cycle consistency loss

To make the image generated by the generator retain more structural information of the input image, and further prevent the occurrence of facial distortion of the stylized image. In this paper, the cyclic consistency loss is applied between the reconstructed images I_c^{rec} and I_s^{rec} generated by the generator $G_{S\to T}$ and their corresponding original input images $I_c \in S$ and $I_s \in T$, which is expressed as follows:

$$I_{cyc} = U^2E_{I_c,I_s}[\|I_c,I_c^{rec}\|_2 + \|I_s,I_s^{rec}\|_2] \tag{14}$$

(4) Overall objective function

Our network model structure target loss includes four parts: local style loss, adversarial loss, cycle consistency loss, portrait reconstruction loss, and the formula is expressed as follow:

$$L = L_{adv} + L_{cyc} + \alpha L_{local\_style} \tag{15}$$

Here, α is the hyperparameter of L_{local_style}, which controls the relative balance between detail preservation and local style loss.

4. Experiment

4.1 Experimental setup

(1) Dataset. 1. Dataset. In our experiments, the content image dataset is from the first 1000 images of the publicly available HD face dataset FFHQ[19]. Style Images contains four datasets: Cartoon, Fantasy, Illustration, and Impasto, all at 1024×1024 resolution. Our goal is to allow users to use the model in this paper to achieve not only local portrait style transfer but also high-quality overall portrait style transfer with only a few hundred style images. Among them, the data sets of Cartoon, Fantasy, Illustration, and Impasto are all from DualStyleGAN[9] and contain 317 images, 137 images, 156 images, and 120 images respectively. Some example images of the dataset are shown in Fig. 5.

Fig. 5. Example images from the dataset in this paper. The datasets (a) to (e) are Cartoon, Fantasy, Illustration, Impasto and Content images, respectively.

(2) Implementation details: We implemented the high-resolution local portrait style transfer network on a single NVIDIA Tesla A100 GPU using PyTorch. It is trained for 25,000, 22,000, 18,000, and 16,000 iterations on the cartoon, abstract, illustration, and thick-painted datasets, respectively. Training each dataset took an average of 8 hours at 1024×1024 resolution with a batch size of 8 and required a memory size of 18GB. In the final total loss formula in Equation (15), the parameter α is 0.08, 0.07, 0.06, and 0.06 for Cartoon, Fantasy, Illustration, and Impasto datasets, respectively, and the learning rate is set to 0.0001.

(3) Evaluation metrics: We selected three objective metrics to evaluate the effectiveness of the model: FID, SSIM, and LPIPS. FID uses common metrics for evaluating the quality of GAN-generated images. SSIM represents the structural similarity index between the original content and the stylized image, which quantifies the level of detail preservation during portrait style transfer, and LPIPS measures the difference between the two images. We tested using these metrics on four data sets separately, and to

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

ensure the comparability of the results of different methods, the same data set and resolution are used for all the comparative experimental methods that will be included. The parameter Settings for each comparison experiment were generated according to the official implementation and default configuration provided by the authors, and all methods were executed on a single NVIDIA Tesla A100 GPU.

(4) Comparison methods: We chose to compare with six state-of-the-art portrait style transfer methods: CycleGAN[13], a general image translation framework; A multi-layer Transformer decoder is used to stylize the content sequence according to the style sequence, StyTr2[20]. Through the fusion of pure content and style features, the expression between content and style is strengthened to ensure the delicacy and consistency of the transfer effect. Puff-Net[21] Small Sample Portrait Style Transfer Method JoJoGAN[22]; This paper designs a Portrait style transfer method DualStyleGAN[9] with a three-stage progressive fine-tuning strategy, and an Unpaired cyclic mapping portrait generation method Unpaired portrait-drawing[23].

4.2 Ablation experiment

To further verify the effectiveness of each component in our model for network performance improvement, four ablation experiments are conducted on the network architecture: U-shaped encoder, generator structure, local portrait style transfer module, and loss function.

(1) Ablation with the U-shaped encoder. Different from the 1×1 or 3×3 convolution kernel size commonly used in traditional CNN[15] or Resnet[19] design, in order to more fully learn the local features of the input image, we design a unique U-shaped encoder, namely content encoder U^2E_c and style encoder U^2E_s, respectively. Its nested U-shaped structure can extract more structural features of the image from different scales, which not only ensures the quality of local portrait style transfer but also improves the overall stylization effect of portrait style transfer. As shown in Fig. 6 (b), when the U-shaped encoder is used, the quality of the stylized image generated by the model is significantly better than the result of traditional portrait style transfer. As shown in Fig. 6 (a), the details of the stylized image are significantly improved, such as the hair color in the first row is uneven. Obvious artifacts in the hair part of the second row, blurred facial lines in the third and fourth rows, etc.

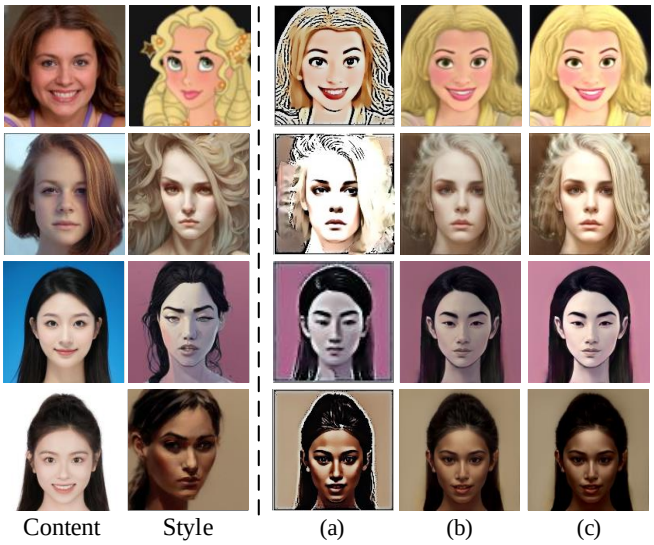

Fig.6. Ablation results for key network components. (a) Result of the traditional portrait style transfer. (b) Result of portrait style transfer using U-shaped encoder. (c) Result of structural portrait style transfer from the optimized generator

(2) Ablation with the generator structure. Unlike traditional portrait style transfer architectures, the input of the generator architecture in this chapter is two-input, and the generator contains two U-shaped encoders. As shown in Fig. 6(c), when the improved generator structure is used, the overall portrait style transfer performance of the model is improved. Compared with Fig. 6(b), the color of the optimized generator structure is closer to the target style image, and more details of the original content image are retained in the structure.

(3) Ablation with the local portrait style transfer module. A very key place for local portrait style

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

transfer is to deal with the area with portrait style transfer and the area without portrait style transfer, only the portrait style transfer is performed on the specific area, and the rest of the area keeps the structure unchanged. As shown in Fig. 7, when using local portrait style transfer, a very important point is the handling of feature fusion. Different from the traditional feature map fusion, we no longer use the traditional feature map splicing way and increase the accuracy of local portrait style transfer by using the style attention mechanism, which fuses the content and style features of the specified region to achieve higher portrait style transfer quality. As shown in Fig. 7(a), when the local portrait style transfer network is not used, although the network learns the style features of the target style image, there are some problems such as facial structure distortion and artifacts of portrait style transfer in specific regions. As shown in Fig. 7(b), after adding the local portrait style transfer module, the local portrait style transfer of the specified region image can accurately learn the style features of the target style image, and solve the problem of facial structure distortion, but there are still artifacts. Fig. 7(c) shows the addition of the Local Feature fusion module (LFFM). The artifacts of the stylized images are significantly reduced and the structural features of the content images are preserved.

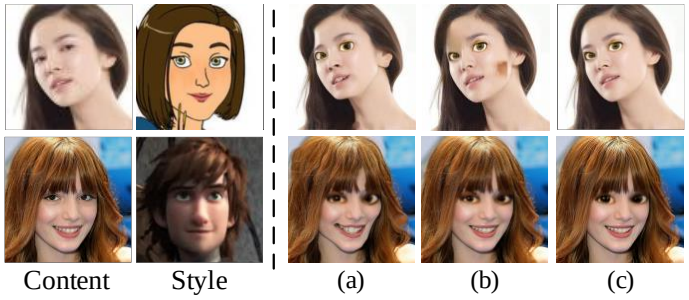

Fig.7. Local portrait style transfer module ablation. (a) Shows the result without using the local portrait style transfer module. (b) Shows the result using local portrait style transfer. (c) Represents the result of (b) +LFFM.

(4) Ablation with the loss function. To verify the effectiveness of the key loss in the proposed model, we divide the overall objective loss function ablation of the model into two versions. Specifically, the local portrait style loss L_{local_style} and cycle consistency loss L_{cyc} are removed respectively to verify the effects of local portrait style transfer contour preservation and target learning style on network performance.

1) Local portrait style loss L_{local_style}: In the case of removing L_{local_style} and only retaining L_{cyc} and L_{adv}, as shown in Fig. 8(a), the generated stylized image does not fully learn the target, including color rendering errors, unsmooth lines, unclear local textures, and other problems, and also affects the integrity of the local portrait facial structure, such as facial distortion. Loss of eyebrows, etc.

2) Cycle consistency loss L_{cyc}: In the case of removing L_{cyc}, as shown in Fig. 8(b), the generated stylized images suffer from severe contour deformation and artifacts, and the integrity and correctness of the basic semantic structure of the portrait face are lost, so the characters in the synthetic images lose a lot of content structure. Finally, in Fig. 8(c), after adding all losses to the network architecture, the overall texture structure of the image is clearer and the local portrait stylization effect is also optimized, indicating that both L_{local_style} and L_{cyc} help to improve the quality of portrait style transfer.

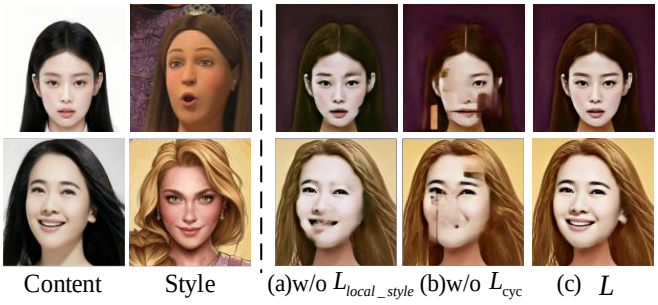

Fig. 8. Loss function ablation results. (a)The result of removing theL_{local_style}. (b)The result of removing L_{cyc}. (c) The result of the full loss function L.

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

4.3 Comparison with state-of-the-art methods

As shown in Fig. 9, this chapter performs overall stylizing experiments on four different artistic styles, demonstrating that our method can effectively preserve the structural features of the original content image while generating globally high-quality stylized images. Comparison methods include: CycleGAN[13], StyTr2[20], Puff-Net[21], JoJoGAN[22], DualstyleGAN[9], and Unpaired Portrait-Drawing[23].

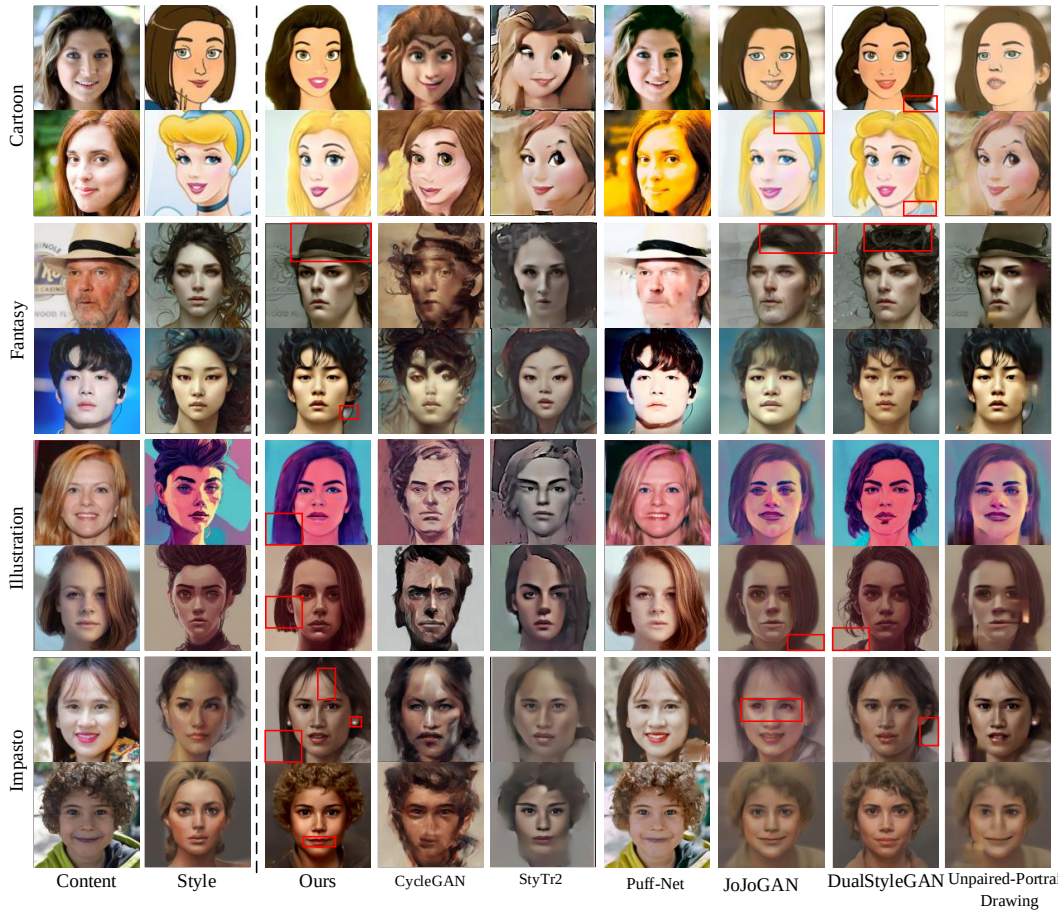

Fig. 9. Results compared to state-of-the-art methods.

The comparison results are shown in Fig. 9, which shows that the stylized images generated by our method can learn the features of the stylized images more fully. Besides our method, DualstyleGAN[9] performs best with JoJoGAN[22]. Although DualstyleGAN[9] can accurately imitate the target style, it cannot accurately preserve the structural features of the original content image, resulting in the loss of important features of the content image in the stylized image, for example, the loss of the man's hat in the third row, and because its method focuses too much on the style features of the style image. This leads to some structural features in the stylized image that are not present in the original content image, such as the shoulders in the second and sixth rows. JoJoGAN[22] pays too much attention to style features, which leads to the phenomenon of over-fitting in stylized images. For example, women's hair accessories appear in the second row, and men's hats disappear in the third. Although other methods StyTr2[20] and Unpaired Portrait-Drawing[23] can learn the target style features, the generated stylized images have obvious artifacts and facial structure distortions, which seriously affect the visual effects of the stylized images. Puff-Net[21] can fully preserve the features of content images, but it does not learn the features of style images, resulting in unobvious style features. Fig. 9 illustrates that our approach yields satisfactory results in various portrait style transfer.

To further show the superiority of our method, SSIM, LPIPS, and FID indicators are used to evaluate the method in this chapter and compare it with other methods in the overall stylization experiment. As shown in Table 1, the stylized images produced by our method have the lowest FID and LPIPS scores, and the highest SSIM values, indicating that the stylized images generated by our method are of good quality, which further verifies that the method in this chapter is superior to other methods.

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

Table 1 Quantitative comparison (The evaluation is performed according to four different styles)

Methods	Cartoon			Fantasy			Illustration			Impasto		
	SSIM↑	LPIPS↓	FID↓	SSIM↑	LPIPS↓	FID↓	SSIM↑	LPIPS↓	FID↓	SSIM↑	LPIPS↓	FID↓
Ours	**0.708**	**0.331**	**149.923**	**0.708**	**0.346**	**149.605**	**0.717**	**0.542**	**141.779**	**0.757**	**0.452**	**112.086**
CycleGAN	0.646	0.411	211.157	0.607	0.37	177.225	0.567	0.582	178.354	0.599	0.628	157.179
StyTr2	0.522	0.493	174.257	0.579	0.558	171.506	0.522	0.493	174.257	0.579	0.558	171.506
Puff-Net	0.551	0.564	187.577	0.601	0.585	179.294	0.649	0.576	188.263	0.575	0.592	161.175
JoJoGAN	0.59	0.767	167.375	0.631	0.638	176.129	0.649	0.56	258.653	0.75	0.663	169.985
DualStyleGAN	0.587	0.408	177.221	0.624	0.661	168.441	0.638	0.656	177.02	0.676	0.661	146.037
Unpaired-Portrait-Drawing	0.562	0.468	203.195	0.583	0.589	155.117	0.573	0.624	169.065	0.512	0.585	205.168

To further verify the effectiveness of the method in this chapter, here is an example of a controlled local portrait style transfer to a specified area. Fig. 10 shows the local portrait style transfer results of different regions for the same content image, where (a) represents the portrait style transfer results of hair and face, (b) represents the portrait style transfer results of eyebrows region based on (a), and (c) represents the portrait style transfer results of eyes region based on (b). (d) represents the result of the portrait style transfer performed based on (c) concerning the mouth region, and (e) represents the result of the portrait style transfer performed based on (d) concerning the nose region. The model in this chapter performs well in local portrait style transfer of different regions in the same content image, and there is no facial structure distortion, artifacts, or other problems that affect visual perception.

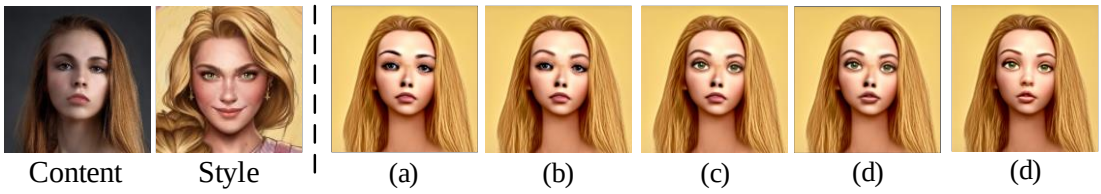

Fig. 10. Local portrait style transfer results for different regions of the same content image.

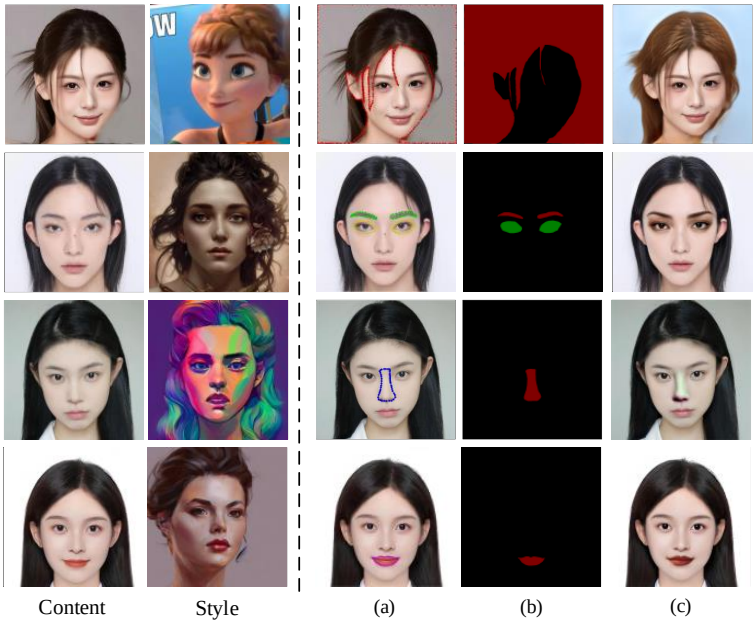

Fig. 11. Local portrait style transfer results for images with different content. (a) Represents the region where the local portrait style transfer is to be performed. (b) A binary mask representing the region to perform local portrait style transfer. (c) Represents the result of the local portrait style transfer region.

Fig. 11 shows the process of local portrait style transfer in different regions for different content images. The first column is the content image, the second column is the style image, the third column is the region that requires local portrait style transfer, the fourth column is the binary mask map of the region, and the fifth column represents the result after local style portrait transfer. The first row is an example of performing local portrait style transfer except for the face, the second row is an example of local portrait style transfer for eyebrows and eyes, the third row is an example of local portrait style transfer for the nose, and the fourth row is an example of local portrait style transfer for the mouth. From

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

the local portrait style transfer results in Fig.10 and Fig. 11, it can be seen that the algorithm in this chapter performs well in the local portrait style transfer, and the method in this chapter can accurately capture the target style of different regions, and realize the local portrait style transfer for arbitrary content images. The stylized results are consistent with the structure of the content image, the texture transitions are natural, and there are no problems such as facial distortions, artifacts, and structural loss that affect the quality of the stylization.

4.4 User study

We conduct a user study to evaluate the proposed method against state-of-the-art style transfer methods: CycleGAN[13], StyTr2[20], Puff-Net[21], JoJoGAN[22], DualstyleGAN[9], and Unpaired Portrait-Drawing[23]. Specifically, this paper sought 35 participants to rank and rate the experimental results of the above six comparison methods and the four artistic styles in this paper. This paper calculates the final average score for each method based on the rankings and scores of the 35 participants, and statistics on the proportion of each method that is considered first. As shown in Table 2, for each artistic style, the majority of participants considered the method in this paper to work best.

Table 2 User study

Styles	Methods						
	Ours	CycleGAN	StyTr2	Puff-Net	JoJoGAN	DualStyleGAN	Unpaired-Portrait-Drawing
Cartoon	80.00%	0.00%	2.86%	2.86%	5.71%	8.57%	2.86%
Fantasy	74.29%	0.00%	0.00%	2.86%	2.86%	19.17%	8.57%
Illustration	77.14%	2.86%	5.56%	0.00%	8.57%	5.71%	8.57%
Impasto	71.43%	0.00%	2.86%	0.00%	5.71%	8.57%	8.57%
Average	75.72%	0.72%	2.82%	1.43%	5.71%	10.51%	7.14%

4.5 Limitations

In Fig. 12, we show two typical cases where the stylization effect is not ideal. Firstly, although the structural features of the content image are well preserved and the style of the target style image is fully learned, when the face in Fig. 12 (a) is occluded, the overall portrait style transfer will lead to local transfer failure, resulting in deformation of the local facial contour. Since the proposed segmentation method cannot separate the foreground and background well, the background part cannot be well preserved during the overall portrait style transfer, as shown in Fig. 12 (b).

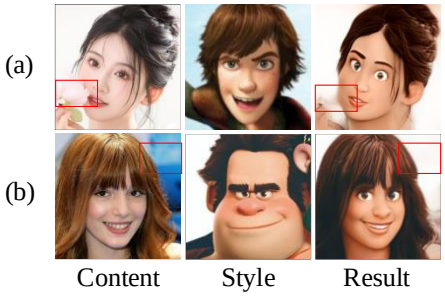

Fig. 12. Limitations of our approach.

5. Conclusion

In this paper, we propose a high-resolution controllable local portrait style transfer model. Different from the traditional holistic image translation framework, this paper introduces a new method to balance the information difference between the content domain and the style domain, so that it can better preserve the parts that have not been transferred when performing partial portrait style transfer, and this method is still applicable to the holistic portrait style transfer. By introducing the encoder in the generator architecture G into the U^2-Net framework, the structural features of the target domain and the style domain can be more fully learned. In addition, a new local portrait style transfer module is proposed, which can not only make the area undergoing style transfer accurately learn the target style, but also maintain the original structural features of the part without style transfer, and generating high-quality stylized images while preserving the structural features of the content image. The Local Feature Fusion Module (LFFM) was designed to further improve the quality of style transfer. Finally, to reduce the

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

artifacts and local contour deformation in the stylized image, histogram matching is used for the area without style transfer, and the local portrait style loss L_{local_style} and cycle consistency loss L_{cyc} are used to constrain the part of portrait style transfer, to reduce the generation of artifacts in local portrait style transfer. And it constrains the contour structure of the style image. In addition, a series of comparative experiments and ablation experiments on four different style datasets show that the proposed method performs well in terms of overall portrait style transfer and local portrait style transfer, which proves the effectiveness of the proposed method in portrait style transfer.

References

[1] Rosin P L, Lai Y-K. Non-photorealistic rendering of portraits[C]//Proceedings of the workshop on Computational Aesthetics. Goslar, DEU: Eurographics Association, 2015: 159-170.

[2] Li B, Zhu Y, Wang Y, et al. AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation[J]. IEEE Transactions on Multimedia, 2022, 24: 4077-4091.

[3] Fišer J, Jamriška O, Simons D, et al. Example-based synthesis of stylized facial animations[J]. ACM Trans. Graph., 2017, 36(4): 155:1-155:11.

[4] Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2021: 852-863.

[5] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4401-4410.

[6] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.

[7] Kong F, Pu Y, Lee I, et al. Unpaired Artistic Portrait Style Transfer via Asymmetric Double-Stream GAN[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9):5427-5439.

[8] Song G, Luo L, Liu J, et al. AgileGAN: stylizing portraits by inversion-consistent transfer learning[J]. ACM Trans. Graph., 2021, 40(4): 117:1-117:13.

[9] Yang S, Jiang L, Liu Z, et al. Pastiche master: Exemplar-based high-resolution portrait style transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7693-7702.

[10] Khowaja S A, Mujtaba G, Yoon J, et al. Face-past: Facial pose awareness and style transfer networks[J]. arXiv preprint arXiv:2307.09020, 2023, 1.

[11] Qi T, Fang S, Wu Y, et al. Deadiff: An efficient stylization diffusion model with disentangled representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8693-8702.

[12] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[13] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[14] Chen J, Liu G, Chen X. AnimeGAN: A Novel Lightweight GAN for Photo Animation[C]. Li K, Li W, Wang H, et al., eds.//Artificial Intelligence Algorithms and Applications. Singapore: Springer, 2020: 242-256.

[15] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2414-2423.

[16] Xing Y, Li J, Dai T, et al. Portrait-aware artistic style transfer[C]//2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 2117-2121.

[17] Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern Recognition, 2020, 106: 107404.

[18] Zhang Q L, Yang Y B. Sa-net: Shuffle attention for deep convolutional neural networks[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 2235-2239.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] Deng Y, Tang F, Dong W, et al. Stytr2: Image style transfer with transformers[C]//Proceedings of

ISSN 2616-5775 Vol. 8, Issue 4: 19-32, DOI: 10.25236/AJCIS.2025.080403

the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11326-11336.
[21] Zheng S, Gao P, Zhou P, et al. Puff-net: Efficient style transfer with pure content and style feature fusion network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 8059-8068.
[22] Chong M J, Forsyth D. Jojogan: One shot face stylization[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 128-152.
[23] Yi R, Liu Y J, Lai Y K, et al. Unpaired portrait drawing generation via asymmetric cycle mapping[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8217-8225.