Displays 80 (2023) 102530

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

WCGAN: Robust portrait watercolorization with adaptive hierarchical localized constraints☆

Hongjin Lyu *, Paul L. Rosin, Yu-Kun Lai

School of Computer Science and Informatics, Cardiff University, Abacws, Senghennydd Road, Cardiff, CF24 4AG, United Kingdom

ARTICLE INFO

Keywords:
Image/video style transfer
Watercolor
Portrait
Local details
Scale adaptive
Temporal consistency

ABSTRACT

Deep learning has enabled image style transfer to make great strides forward. However, unlike many other styles, transferring the watercolor style to portraits is significantly challenging in image synthesis and style transfer. Pixel-correlation-based methods do not produce satisfactory watercolors. This is because portrait watercolors exhibit the sophisticated fusion of various painting techniques in local areas, which poses a problem for convolutional neural networks to accurately handle fine-grained features. Moreover, the common but problematic way of coping with multiple scales greatly impedes the performance of existing style transfer methods with fixed receptive fields. Although it is possible to develop an image processing pipeline mimicking various watercolor effects, such algorithms are slow and fragile, especially for inputs of different scales. As a remedy, this paper proposes WCGAN, a generative adversarial network (GAN) architecture dedicated to watercolorization of portraits. Specifically, a novel localized style loss suitable for watercolorization is proposed to deal with local details. To handle portraits of different scales and improve robustness, a novel discriminator architecture with three parallel branches of varying sizes of receptive fields is introduced. In addition, the application of WCGAN is expanded to video style transfer where a novel kind of video training data based on random crops is developed to efficiently capture temporal consistency. Extensive experimental results from qualitative and quantitative analyses demonstrate that WCGAN generates state-of-the-art, high quality watercolors from portraits.

1. Introduction

Watercolor paintings with various distinctive effects are made by delicately controlling the distribution of water and pigments. However, due to its complexity, even artists with long-term professional training need to spend enormous time and effort to complete high-quality watercolor paintings, not to mention ordinary people.

Many works [1–3] in the field of non-photorealistic rendering (NPR) have studied how to transfer images into different styles such as sketch, paper-cut and oil painting. In particular, Rosin and Lai [4] developed a specific image processing pipeline that tries to mimic different effects of watercolor. The method achieved high-quality watercolor stylization of portraits. However, its slow run-time seriously hinders its application, and the method may fail to produce good results for challenging input.

Gatys et al. [5] discovered that the features extracted from convolutional neural networks (CNNs) can characterize visual styles. The subsequent works have spent considerable effort to enhance the style transfer performance from different perspectives, which can be roughly divided into: generic and specific style transfer.

Generic style transfer works [6–10] can be categorized into three directions: minimizing specific measures for content and style dissimilarities, aligning feature distributions between the content and style images, and learning to transfer between different domains. While these methods have made great progress, the trade-off between generalization and quality limits their performance in portrait watercolorization.

A few works [11–14] center on simulating specific style characteristics, where local fusion and multi-scale inputs are general problems. These components play a particularly important role in the portrait watercolorization task. The fluidity of water and the transparency of pigment allow watercolors to display extraordinary beauty, which is the result of sophisticated fusion of multiple effects (wobbling, diffusion, edge darkening, etc.). This makes the quality of local multi-effect fusion directly affect the aesthetic feeling of watercolor. Moreover, faces may be of varying sizes in the image, e.g. depending on the distance to the camera, which requires the watercolor portrait to pay extra attention to multi-scale input processing. These high requirements make portrait watercolorization challenging for existing neural style transfer methods.

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author.
E-mail addresses: lyuh2@cardiff.ac.uk (H. Lyu), rosinpl@cardiff.ac.uk (P.L. Rosin), laiy4@cardiff.ac.uk (Y.-K. Lai).

https://doi.org/10.1016/j.displa.2023.102530
Received 31 May 2023; Received in revised form 1 August 2023; Accepted 28 August 2023
Available online 1 September 2023

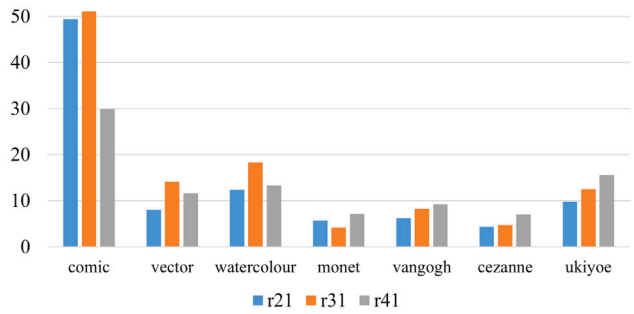
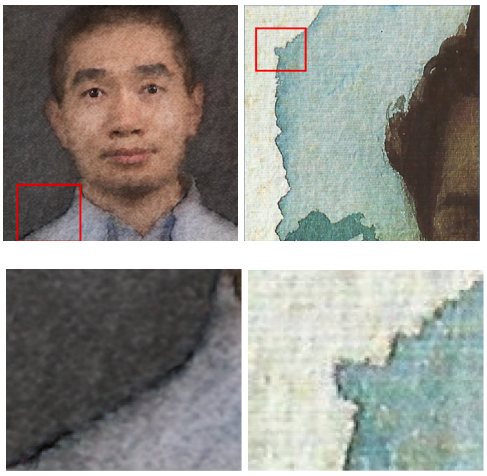


Fig. 1. Style variation comparison.



(a) WCGAN results (b) Professional watercolor

Fig. 2. Comparison with self-portrait by Enrique Campo Sobrino (1890–1911).

To demonstrate the challenges of watercolorization, we quantify the style variation (*SV*) of style datasets (*StyleD*), where a larger *SV* indicates a more inconsistent style. Seven different real style datasets are collected from [11,15]. To accurately measure the style inconsistencies, the Gram matrix of pre-trained VGG-19's feature maps, widely employed in the field of image style transfer [5], is utilized to represent style features. Then, the *SV* within *StyleD* is characterized by the trace of the covariance matrix of all style features. As shown in Fig. 1, blue, orange, and gray represent the *SV* results corresponding to the r21, r31, and r41 layers, respectively. The *SV* values of watercolor rank 2nd, 2nd, and 3rd in the r21, r31 and r41 layers, respectively, which indicates that watercolor paintings exhibit a high degree of stylistic variation. Note that the only style with noticeably higher SV is comics. However, this is understandable as comics in the database are highly varied, from black-and-white line drawings to color comics.

To the best of our knowledge, there are no neural style transfer works specifically for portrait watercolorization. Although generic style transfer methods can perform portrait watercolorization, due to the complexity of the watercolor effects, existing general purpose methods cannot achieve satisfactory performance. This is particularly crucial for portraits where even minor defects can be detrimental.

Considering the aforementioned problems, this paper proposes WCGAN, a novel GAN-based [16] approach, to transform portraits to watercolor while preserving the original content and performing watercolor effects. As shown in Fig. 2, the generated results of WCGAN and a professional watercolor painting both exhibit typical watercolor characteristics, namely edge-darkening, wobbling, diffusion and the unique texture, which effectively demonstrates the effectiveness of WCGAN. Our main contributions are summarized below:

- We propose a novel block-based loss term named localized Gram Matrix loss (LGML), which provides an extra fine-grained constraint in local areas, significantly boosting local stylization performance.
- We develop a new Adaptive Discriminator architecture (ADA_dis), which better preserves original image information at different scales. Thanks to this architecture and a multi-scale training dataset, we can adequately handle portraits of different sizes.
- This paper expands WCGAN to video style transfer tasks, where a novel method for generating video training data based on still images can help our network effectively achieve temporal consistency.

The structure of the paper is described as follows: Section 2 introduces the current status of existing style transfer research, especially highlighting research closely related to ours. Section 3 provides corresponding implementation details of WCGAN. Section 4 validates our design by comparison with state-of-the-art networks. The conclusions and future work are in Section 5.

2. Related work

2.1. Non-photorealistic rendering

Non-photorealistic rendering (NPR) focuses on the computer generation of various styles, where a wide variety of stylizations have been applied to portraits in NPR. For example, [1–4] focus on sketch, papercut, watercolor and oil painting, respectively. Traditional NPR methods can produce high-quality stylized results based on accurate feature simulation, but they are often slow. In particular, such simulation of real-world process often leads to a complex pipeline, which can be less robust. Neural Style Transfer (NST), especially the methods based on feedforward networks, is faster but the quality is largely dependent on the training data. In this paper, we propose a GAN based method that combines the strengths of NPR and NST where the NPR method [4] is applied to produce a high quality training dataset. This enables the complicated watercolor appearance to be properly and quickly depicted.

2.2. Neural style transfer

The pioneering work [5] formulated the NST task as generating an image that optimizes both a style and a content loss. There are many follow-up works that improve NST both in quality and efficiency, which can be divided into generic and specific style transfer.

Generic Style Transfer aims to transfer multiple arbitrary artistic styles using the same architecture, which can be divided into three strategies:

(1) Perceptual Loss Optimization: These methods aim to produce an output image that minimizes the content loss w.r.t. the content image and style loss w.r.t. to the style image. Follow-up works have further improved this strategy, including using a feedforward network for real-time generation and better handling multiple styles [6,8,9]. Taking a holistic view for the style loss handles lack of correspondence between the output and style image. However, it cannot well capture spatial variation of styles, which limits their application in portrait watercolorization. As a remedy, we propose localized Gram Matrix loss (LGML) specifically focusing on the fine-grained style pattern.

(2) Feature Distribution Alignment: This category conducts the alignment process between the feature distributions of content and style images. The works [17–20] generate stylized images by matching the mean/variance, whitening/coloring feature transforms, relaxed cross-correlation and manifold alignment, respectively. But the pre-trained networks in [19] are trained by normal (non-stylized) images, which cannot achieve satisfactory results in style transfer tasks. Moreover, scale-adaptivity in [20] is achieved through their hourglass network,

which may not be easily generalized to other style transfer studies. In this paper, the proposed ADA_dis enables WCGAN to be scale-adaptive.

(3) Domain Transfer: Other research like GAN-based methods addresses the style problem as transferring between two (or more) domains. Among the numerous applications of GANs [21–25], Pix2Pix [7] develops a generic framework for achieving paired image translation tasks. CycleGAN [10] introduces the cycle consistency losses to deal with unpaired image translation tasks. However, neither have sufficient flexibility to learn spatially varying style features due to the lack of a specific learning mechanism for regional features. Facing the above problems, additional masks and discriminators for corresponding areas have been added in WCGAN which enhance the flexibility of style feature learning. Recently, methods [26,27] based on probabilistic diffusion models have become increasing popular for style transfer. Although such methods show excellent performance when there are a large number of training examples, the performance may drop with limited training data.

In addition, generic style transfer papers show some impressive stylization results, and then make an implicit assumption that their methods can also achieve good performance on any other styles. However, in practice these methods cannot achieve acceptable performance for all styles, so some research considers developing dedicated methods for certain styles.

Specific Style Transfer: A few works have been designed for specific style transfer tasks. PairedCycleGAN [11] and Beauty-GAN [13] focus on transferring the makeup style by two asymmetric networks. CartoonGAN [12] proposes two specific losses for cartoon style: high-level semantic loss and edge loss, respectively. APDrawingGAN [14] applies an extra distance transform loss which focuses on stroke lines. The work [28] further extends it to learn from unpaired training data, by introducing an asymmetric cycle consistency loss to cope with the substantial information gap between photos and line drawings. Similarly, portrait watercolor contains unique effects that are challenging for existing NST works, as discussed in Section 1. Considering the above limitations, WCGAN is proposed specifically for the portrait watercolorization task, although key ideas developed can also be generalized to other challenging style transfer tasks.

2.3. Video style transfer

Video style transfer differs from image style transfer in temporal consistency. Ruder et al. [29] attempt to capture this by applying temporal losses guided by optical flow, but their optimization-based method is time-consuming. Chen et al. [30] achieve long-range consistency via a recurrent neural network architecture, which requires slow optical flow calculation in the inference stage. The stylization process of [31] is faster than [29,30] which require optical flow calculation, but [31] only calculates content loss based on one layer (relu4-2), which cannot capture subtle textures and strokes. Gao et al. [32] improve the temporal stability by adding an extra luminance constraint. However, fine-grained texture features are not captured due to the lack of local style constraints.

Moreover, there are no existing video training datasets available for portrait watercolorization. This paper proposes a new method to generate video training data using still images in a self-supervised manner, which can meet the requirements for availability and accuracy of the video training dataset.

3. Watercolor transfer of portrait photography

This paper proposes a GAN framework specifically for portrait watercolorization which contains a generator G and a discriminator D. We regard the process of transferring a portrait photograph in domain $\mathcal{P}$ into a watercolor painting in domain $\mathcal{W}$, whilst preserving the content of the original portrait photograph, as a mapping function. This watercolorization mapping function is learned from the paired training dataset $T_{training} = \{(p_i, w_i) \,|\, p_i \in \mathcal{P}, w_i \in \mathcal{W}, i = 1, 2, \ldots, N\}$, where N is the total number of portrait-watercolor pairs and i is the index number. Denote L as the overall loss function, which contains three terms: L_{L1}, L_{LGM} and L_{adv}, corresponding to pixel-wise $L1$ loss for content preservation, local Gram matrix loss for style preservation, and adversarial loss. WCGAN is trained by solving the following min–max problem:

$$\min_G \max_D L(\lambda_1 L_{L1}(G) + \lambda_2 L_{LGM}(G) + \lambda_3 L_{adv}(G, D)), \tag{1}$$

where λ_1, λ_2 and λ_3 are weights that balance the importance of the loss terms.

In the following sections, this paper introduces the detailed architecture of the Generator and Discriminator in Sections 3.1 and 3.2 respectively. As shown in Fig. 3, both of the Generator and Discriminator are fully convolutional, which means the same network can learn to handle input images of different resolutions. The hybrid loss function L is described in Section 3.3. Moreover, a new way of generating video training data using still images for temporal consistency is introduced in Section 3.4.

3.1. Generator

The aim of G is to render a portrait photograph in a watercolor style, while keeping the content structure of the original portraits. To capture multiple abstraction degrees meaningful for watercolor portraits, we generate 7 semantic masks (M^*, where * refers to individual regions) in advance and directly add them to the input of G, which enables G to learn multiple independent features. The seven corresponding areas are: eyebrows, eyes, nose, inner-mouth, outer-mouth, face, skin as shown in Fig. 4. Based on OpenFace [33], the first five masks accurately indicate specific facial regions. The face mask refers to the entire face region in the OpenFace results. Additionally, the Skin mask serves as a supplement to the OpenFace results, providing additional skin information such as the neck and ears. Fig. 3(a) shows the structure of the Generator, which is a traditional Encoder-Decoder structure. The encoder part consists of 8 down-sampling convolution layers with stride 2 and 4×4 kernels. The desired watercolor paintings are reconstructed after 8 up-sampling convolution layers with the same stride 2 and 4×4 kernels. Eight skip connections between the encoder and decoder can effectively recover fine-grained details.

3.2. Discriminator

3.2.1. Hierarchical discriminator architecture

A hierarchical Discriminator structure D is proposed in WCGAN, following [14]. D returns multiple scores corresponding to the different regions used in G, which provides a more comprehensive judgment compared with returning only one score. This method is also in line with artists who adopt different drawing techniques for different parts. $D = \{D_{global}, D_{\mathcal{L}}\}$, where D_{global} judges whether the input is real or fake based on the whole image. $D_{\mathcal{L}}$ consists of 7 separate discriminators which focus on the performance of the local facial regions listed in Section 3.1.

3.2.2. Adaptive discriminator architecture

The multi-scale problem in image processing is common as faces may vary in size depending on the distance to the camera. Obviously, a neural network with only one fixed receptive field cannot accurately recognize features at different scales. Dilated convolution [34], can change the receptive field while still keeping the total number of parameters unchanged. Based on these observations, we propose a novel adaptive discriminator architecture with three parallel branches as shown in Fig. 5, which have different dilation rates enabling different receptive fields. In addition, different regions of the same image may require different receptive fields in real-life scenarios. For example, an image may contain background objects at a distance (so would benefit

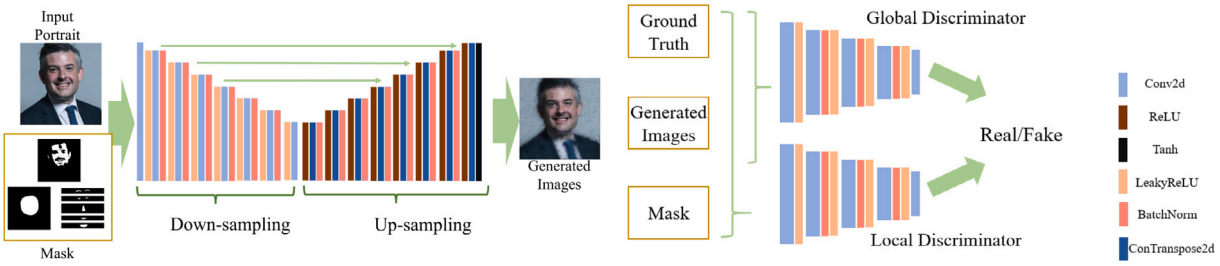


Fig. 3. Overview of the WCGAN architecture.

Fig. 4. Seven masks corresponding to semantic regions important for portrait stylization.

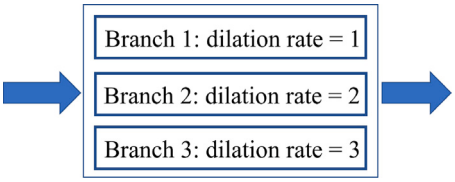


Fig. 5. Adaptive discriminator architecture.

Table 1
Time consumption for [4], Updated [4] and WCGAN in Section 4.1.

Resolution	[4] (s)	Updated [4] (s)	WCGAN (s)
256^2	50.37	119.60	**0.04**
512^2	82.10	120.21	**0.15**
1024^2	213.03	119.73	**0.62**

from small receptive fields) along with faces which are much closer (requiring larger receptive fields). Thus, the max-pooling result of the three branches' outputs is regarded as the final output to preserve the strongest responses from different branches, which flexibly enables suitable receptive fields to different regions.

Some existing works [35,36] also proposed multi-scale discriminators that apply three identical discriminators but operate at different image scales, where simple down-sampling operations are used, which lose a lot of important information. Moreover, assigning weights manually to different branches' outputs often requires manual tuning. Although the weights are adjusted using a predetermined formula during training in [35], fixed weights are still used in each training iteration to fuse multi-scale features [35,36], lacking flexibility to handle different image scenarios. Our proposed ADA_dis processes images of the same scale with branches of different scales, and merges features with max-pooling, flexibly assigning suitable receptive fields for different regions and bypassing the issue of weight assignment.

3.3. Loss function

The hybrid loss function L consists of three parts: L_{L1}, L_{adv} and L_{LGM} as shown in Eq. (1). With the help of L, the hierarchical adaptive structure of D can drive G to produce watercolor paintings with a variety of subtle effects and more detailed local area performance.

3.3.1. L1 loss

Least Absolute Deviations ($L1$) is widely used as a loss term in machine learning, which compares the similarity of two pictures from a pixel-wise perspective. According to the following Eq. (2), we calculate the sum of all absolute differences to judge the quality of the generated watercolor painting at the pixel level.

$$L_{L_1}(G,D) = \mathbb{E}_{(p_i,w_i)\sim T_{\text{training}}}\left[\left\|G\left(p_i\right) - w_i\right\|_1\right] \tag{2}$$

3.3.2. L_{adv} adversarial loss

L_{adv} applied in this paper contains L_{global} and L_{local} for global and local discrimination.

L_{global}: helps G approximate the optimal result through a neural network (D_{global}), thereby avoiding the difficult probability calculation problem for generative models. L_{global} is defined as:

$$\begin{aligned} L_{global}(G,D_{global}) = \mathbb{E}_{(p_i,w_i)\sim T_{\text{training}}}\big[&\log\left(D_{global}\left(p_i,w_i\right)\right) \\ &+\log\left(1-D_{global}\left(p_i,G\left(w_i\right)\right)\right)\big] \end{aligned} \tag{3}$$

D_{global}, as the global discriminator, determines the authenticity based on the entire input image at coarse granularity. In real portrait watercolors, certain facial features are much more important, and artists tend to draw them differently. However, only using semantic masks and one global discriminator cannot guarantee multiple abstract degrees to be properly captured, due to the absence of necessary constraints on key areas. Thus, parallel local discriminators $D_{\mathcal{L}}$ are proposed which contains 7 local discriminators each corresponding to a mask region.

L_{local}: as a supplement of D_{global}, $D_{\mathcal{L}}$ focuses on the style transfer quality in the regions specified by M^*, and D_{m} is an individual local discriminator. M^* refers to all the 7 masks. These have the same resolution as the input image. Local discriminators transform the input image into a high-dimensional feature map, denoted as *FM_img*, which is of lower resolution than the input image. M^*_{ds} is the down-sampled version of M^* to match the resolution of *FM_img*. We can then perform element wise multiplication to only retain part of the feature map that is within the masks. L_{local} is defined as:

$$\begin{aligned} L_{local}(G,D_{\mathcal{L}}) = \sum_{D_{\text{m}}\in D_{\mathcal{L}}} \mathbb{E}_{(p_i,w_i)\sim T_{\text{training}}}\big[&\log\left(D_{\text{m}}\left(p_i,w_i\right)\right) \\ &+\log\left(1-D_{\text{m}}\left(p_i,G\left(w_i\right)\right)\right)\big]. \end{aligned} \tag{4}$$

(a) Input (b) [4] (c) Updated[4] (d) WCGAN

Fig. 6. Comparison with the NPR method for watercolor portrait stylization [4] and its updated version to cope with different scales.

3.3.3. L_{LGM} loss

The sophisticated fusion of local effects creates the distinctive beauty of watercolors. However, it is particularly problematic for watercolor stylization where different painting techniques are often applied to individual regions. To address this, we propose a novel loss term L_{LGM} based on a localized Gram matrix to improve the style transform quality in local regions.

Given an input image pair $\mathcal{A}$ and $\mathcal{B}$, to calculate L_{LGM}, we first split them into Z same-sized blocks respectively: $\{(\mathcal{A}_i, \mathcal{B}_i) | \mathcal{A}_i \in \mathcal{A}, \mathcal{B}_i \in \mathcal{B}, i = 1, 2, \ldots, Z\}$. Secondly, the style loss of each pair of corresponding blocks is calculated by the *Styleloss()* function. Finally, the average value of style losses for all corresponding blocks is used as the corresponding L_{LGM} between $\mathcal{A}$ and $\mathcal{B}$. The definition is as follows:

$$L_{\mathrm{LGM}}(\mathcal{A}, \mathcal{B}) = \frac{1}{Z} \sum_{i=1}^{Z} Styleloss(\mathcal{A}_i, \mathcal{B}_i) \tag{5}$$

$$Styleloss(\mathcal{A}_i, \mathcal{B}_i) = \sum_{l \in \{l_s\}} \left\| Gram\left(F^l(\mathcal{A}_i)\right) - Gram\left(F^l(\mathcal{B}_i)\right) \right\|^2 \tag{6}$$

$$Gram(F^l)_{ij} = \sum_k F^l_{ik} F^l_{jk} \tag{7}$$

where F^l is the feature map from layer l of VGG network [37]. $Gram()$ is the Gram matrix calculated based on the corresponding input, where its (i, j) element is essentially the inner product between the vectorized ith and jth feature maps.

Although Gram matrices are commonly applied to entire images as a way to extract texture features, recent works also applied Gram matrices to reflect the texture quality of the region of interest, where the feature maps of region-based methods [38,39] come from the local discriminator, and those of mask-based methods [40] come from semantic masks. However, the above region/mask-based methods just deliver the overall texture quality on the region of interest, which are not fine-grained enough to handle the sophisticated multi-effect fusion. The block-based LGML in this paper provides fine-grained texture quality assessment that can accurately evaluate the quality of multi-effect fusion.

3.4. Video style transfer

To achieve video style transfer, three extra techniques are applied: a novel method for generating video training data called Multi-Crop Video Training data, a temporal loss term and a consecutive-frame-pair training mechanism. The latter two techniques are widely applied in other works, and are explained in detail in the supplementary material.

This paper proposes a novel method for generating video training data by cropping still images. Given a still image, the cropped areas after multiple random cropping operations can be regarded as consecutive video frames. Consequently, the ground truth optical flow between any two cropped areas can be calculated based on the known cropping positions. Based on the above method, a video training dataset containing ground truth optical flow can be established based on only still images, which greatly reduces the difficulty of obtaining video training data. As we will later show by experiments, Multi-Crop video training data can help the network capture temporal consistency. Furthermore, this novel kind of training data does not conflict with the existing training data (based on real video frames). Related experiments verify that the optimal dataset is a collection of video training data generated by both methods.

4. Evaluation

The comparison with traditional NPR work is shown in Section 4.1. The datasets and implementation details are shown in Section 4.2. The comparison with representative paired works is displayed in Section 4.3. Extra analyses about ADA_dis and LGML are shown in Sections 4.4 and 4.5, respectively. The analysis of the contribution of every component is shown in Section 4.6. The evaluation of video style transfer is presented in Section 4.7. Our method can also be extended to other styles with similar characteristics. The experimental results of artistic portrait drawing (APDrawing) style transfer tasks are displayed in the supplementary material.

4.1. Compared with traditional NPR method [4]

The NPR method [4] proposed an image processing pipeline specifically for portrait watercolorization, which struggles with low resolution images. Thus, this paper proposes an updated version of [4] where images are initially resized to make the face a fixed size, and then [4] is applied to generate watercolor results, before resizing them back to the original resolution.

Directly applying [4] or Updated [4] to portrait watercolorization task has clear drawbacks. Firstly, both [4] and Updated [4] require a long processing time. The average time consumption on the test dataset is shown in Table 1, where WCGAN greatly improves the processing speed by more than two orders of magnitude. Secondly, Fig. 6 shows the results of [4], Updated [4] and WCGAN at 256^2. [4] produces an over-stylized effect especially for eyes under low resolution. Updated [4] cannot handle images with multiple faces of varying sizes due to difficulties in achieving different rescalings for each face simultaneously. In contrast, WCGAN can stably generate proper results even for challenging inputs (multiple faces, multiple scales).

4.2. Dataset and implementation details

It is challenging to obtain a valid watercolor training dataset due to the wide variation of style in the real watercolors as shown in Fig. 1. To obtain a watercolor training dataset with consistent style features, [4] is applied to generate watercolor style in portraits. To enhance scale-adaptivity, an adaptive training dataset $T_{training}$, which expands each original training image to three different sizes, is created based on Updated [4].

2000 portrait images with various skin colors and backgrounds are collected from the internet. Each portrait is firstly resized to three resolutions: 256^2, 512^2, and 1024^2. Then, we partition all images into $T_{training}$ of size 1600×3 and T_{test} of 400×3. In addition, 100 photos containing faces of different sizes are collected as a multi-face test dataset $T_{multiface}$, which can better test different methods when handling faces of different sizes. The Adam optimizer is used, where the learning rate = 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$.

4.3. Comparison with state of the art

WCGAN is compared with five of the latest paired data based works: Gatys et al. [5], Pix2Pix [7], I2ICDAE [42], pSp [41] and BBDM [27]. The following content is conducted from three aspects: qualitative analysis (Figs. 7 and 9), quantitative analysis (Tables 2 to 4) and results of user studies (Fig. 10).

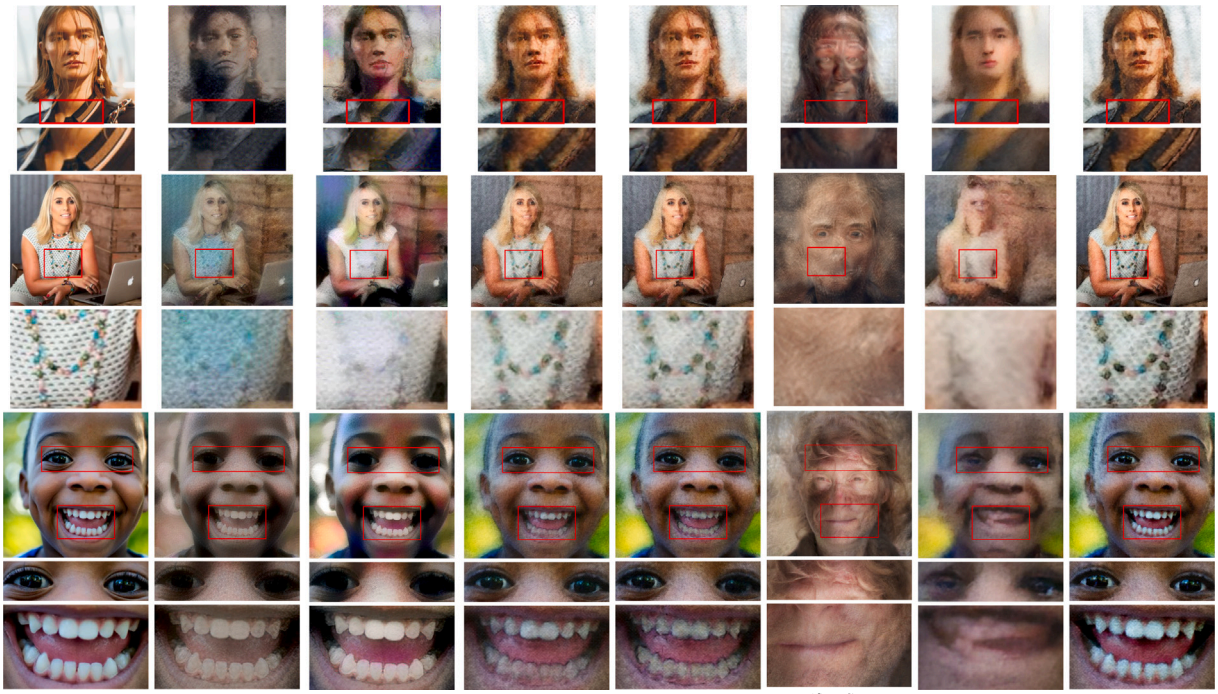


Fig. 7. Comparison with five representative works in Section 4.3.1. 1st–2nd rows: resolution 256^2; 3rd–4th rows: resolution 512^2; 5th–7th rows: resolution 1024^2.

Fig. 8. Comparison with pSp [41] trained and tested on the aligned version of $T_{training}$ and T_{test}.

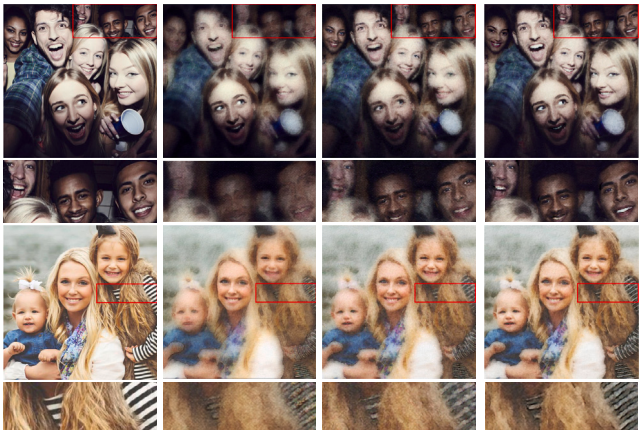


Fig. 9. Comparison with other works on $T_{multiface}$.

4.3.1. Qualitative analysis

Fig. 7 shows the test results of three portraits at different resolutions, where the corresponding close-ups are shown below the full images.

Gatys et al. [5] require as input both a content image and a style image. In our implementation of Original Gatys, for one content image, the watercolor picture in $T_{training}$ with the smallest perceptual loss with this content image is regarded as the corresponding style image. Furthermore, a variant of [5] called Collection Gatys uses the average Gram matrix of all watercolor paintings in $T_{training}$ to measure watercolor style features. The 2nd and 3rd columns of Fig. 7 show the results of Original Gatys and Collection Gatys, respectively. Original Gatys exhibits serious color mismatching at all scales due to its reliance on similar structures between the style and content images, even if the ones with minimum structure difference are picked. Compared with Original Gatys, Collection Gatys improves the performance for details to a certain extent. However, under the same transfer mechanism, Collection Gatys still cannot cope with the color mismatching problem.

I2ICDAE achieves domain transfer by embedding a fully connected layer (FCL) between a pre-trained encoder and decoder, which cannot handle multi-scale inputs. Thus, three independent I2ICDAE models for different input sizes are trained, respectively. The test results of I2ICDAE are shown in the 4th column of Fig. 7. The unacceptable quality of multiple effects fusion affects the overall aesthetics due to the lack of specific loss term or mechanism focusing on local features.

The 5th column in Fig. 7 displays the results of Pix2Pix. The complicated fusion of local effects and independent degrees of abstraction of different regions are not learned properly, which greatly affects the aesthetic feeling, and indicates that Pix2Pix does not have sufficient flexibility to simulate watercolor style.

The 6th column in Fig. 7 displays the results of pSp [41]. Three separate pSp networks are trained for different scales: 256^2, 512^2, and 1024^2. For the 256^2 and 512^2 pSp models, input and output were maintained at the same scale, similar to WCGAN's setup. The 1024-scale pSp model is implemented with the recommended settings by

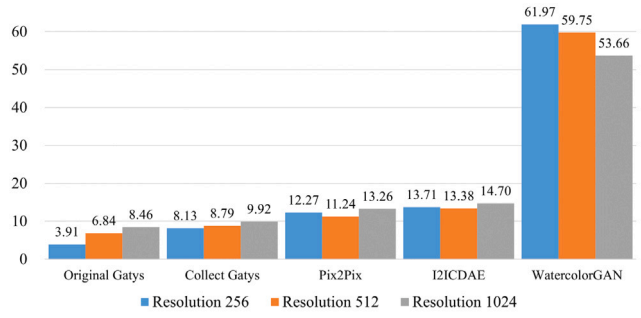


Fig. 10. Stylization preferences captured by the user study in Section 4.3.3.

Table 2
Quantitative comparison with state-of-the-art methods at resolution 256^2 in Section 4.3.2.

Resolution 256 × 256	MSE $\times 10^{-2}$	PSNR dB	SSIM $\times 10^{-2}$	LPIPS $\times 10^{-2}$	FID
Original Gatys [5]	13.32	15.88	42.76	40.21	129.15
Collect. Gatys [5]	7.39	17.83	48.37	34.57	140.33
Pix2Pix [7]	1.37	25.12	74.34	13.78	46.45
I2ICDAE [42]	1.34	25.22	75.65	13.31	44.58
pSp [41]	10.00	16.29	35.52	45.05	150.45
BBDM [27]	3.12	21.33	61.89	31.87	86.76
WCGAN	**1.07**	**26.28**	**81.23**	**11.37**	**40.62**

Table 3
Quantitative comparison with state-of-the-art methods at resolution 512^2 in Section 4.3.2.

Resolution 512 × 512	MSE $\times 10^{-2}$	PSNR dB	SSIM $\times 10^{-2}$	LPIPS $\times 10^{-2}$	FID
Original Gatys [5]	9.21	17.15	49.27	41.32	87.81
Collect. Gatys [5]	5.82	18.89	52.30	37.76	84.62
Pix2Pix [7]	1.11	26.00	73.21	18.33	38.60
I2ICDAE [42]	1.09	26.12	75.55	17.34	36.83
pSp [41]	9.27	16.66	45.98	48.45	223.53
BBDM [27]	2.99	21.52	63.58	40.95	82.43
WCGAN	**0.90**	**26.96**	**78.59**	**15.55**	**33.18**

Table 4
Quantitative comparison with state-of-the-art methods at resolution 1024^2 in Section 4.3.2.

Resolution 1024 × 1024	MSE $\times 10^{-2}$	PSNR dB	SSIM $\times 10^{-2}$	LPIPS $\times 10^{-2}$	FID
Original Gatys [5]	6.36	18.67	52.21	40.55	76.27
Collect. Gatys [5]	5.66	19.04	52.83	40.43	80.92
Pix2Pix [7]	1.07	26.00	69.85	21.13	35.79
I2ICDAE [42]	0.98	26.44	74.29	**18.02**	37.27
pSp [41]	9.81	16.44	49.78	61.84	126.36
BBDM [27]	2.94	21.55	65.42	57.21	97.89
WCGAN	**0.97**	**26.55**	**75.69**	19.47	**33.75**

authors to achieve the proper performance of pSp and alleviate the computational burden of training with 1024^2 input and output (by scaling the input to lower-resolution 256^2 size, and generating output stylized images to 1024^2). The results at 256^2 and 512^2 show that pSp's output cannot present reasonable faces with watercolor style. The results at 1024^2 demonstrates that, despite adhering to the recommended default settings, pSp consistently exhibits undesirable inconsistencies in facial regions compared to the input portraits. The underlying cause is that the pSp model is limited by the pre-trained StyleGAN2 model which is trained on aligned inputs (i.e. the portraits are rotated and cropped to normalize the face orientation and scale), and fails to cope with unaligned inputs present in our watercolor dataset.

Moreover, we further create an aligned version of both $T_{training}$ and T_{test}. By eliminating the confounding factors introduced by unaligned images, we comprehensively validate the performance of pSp. As shown in Fig. 8, the pSp model trained and tested on aligned images still fails to produce results with stable facial consistency compared to the input portraits. Furthermore, it also struggles to accurately simulate watercolor painting textures and complex multi-effect fusion. These findings further corroborate the limitations of the pSp model in achieving high-quality watercolor style transfer.

The 7th column in Fig. 7 displays the results of BBDM (Brownian Bridge Diffusion Model) [27]. To ensure fairness, the official implementation is used, where BBDM can only handle images of 256^2. Thus, each image is initially resized to 256^2 before further processing. For images of 512^2 and 1024^2, the generated images are up-sampled to its original scale. This process may lead to some loss of details. One of the advantages of BBDM is its ability to generate multiple plausible outputs given a single input due to the inherent ambiguity of domain transfer. BBDM generates five slightly different outputs, and the one with the best performance is selected as the final result. BBDM fails to produce results with acceptable quality, in terms of fusion of local effects, locally varying facial abstraction levels, and watercolor texture. The underlying reason is that BBDM as a diffusion based method has the advantage of learning more effectively from a large amount of training samples, but the performance may not be satisfactory when training data is limited.

Since I2ICDAE and Pix2Pix have relatively similar performance to WCGAN, to prove the superiority of WCGAN more convincingly, the generated results of I2ICDAE, Pix2Pix and WCGAN on $T_{multiface}$ are shown in Fig. 9. The first photo contains faces of different sizes. Although the trained I2ICDAE and Pix2Pix can generate tolerable results for simple photos, both methods cannot achieve the desired facial results in complex situations that include multiple faces of different scales. The second photo contains complicated local areas, i.e., continuous small areas with different colors. Due to the lack of a specific mechanism to thoroughly learn the feature of local areas, both I2ICDAE and Pix2Pix cannot present proper effect simulation for the above situation.

4.3.2. Quantitative analysis

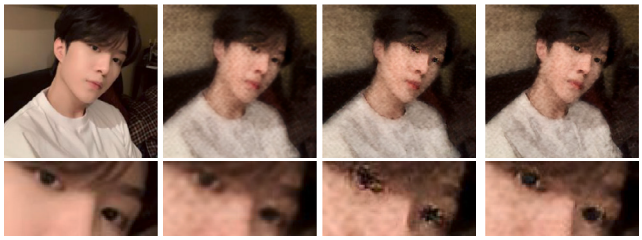
Five common metrics (mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), Fréchet inception distance (FID score) and perceptual metric (LPIPS)) are applied to quantitatively measure all methods on T_{test}. Tables 2–4 respectively show the results of all methods for portraits of different sizes, and the following conclusions can be made:

- Due to the severe color mismatching appearing in both variants of the Gatys method, WCGAN greatly outperforms both Gatys methods under all evaluation criteria.
- Pix2Pix still has a certain gap compared with WCGAN. In particular, the performance of Pix2Pix under the three scales is worse than that of WCGAN 15.91% in LPIPS and 12.24% on FID on average.
- WCGAN achieves an overall improvement under all scales compared with I2ICDAE. The LPIPS score of I2ICDAE at resolution 1024 is slightly better than that of WCGAN. However, it is undoubtedly clear that WCGAN can more precisely simulate watercolor style, whereas each I2ICDAE model only needs to deal with the style transfer task of a single scale.
- Due to the unacceptable issue of facial inconsistency and poor simulation of watercolor painting features in pSp's generated results, the quantitative analysis of pSp experiment results are worse than other methods.
- WCGAN outperforms BBDM by a significant margin in all evaluation metrics, which is consistent with the conclusions drawn from the qualitative analysis.

Overall, WCGAN achieves the best portrait watercolorization on different scales. This is consistent with the aforementioned qualitative analysis results.

Table 5
Quantitative comparison with [35,36].

	256^2		512^2		1024^2	
	LPIPS	FID	LPIPS	FID	LPIPS	FID
[36]	20.20	61.38	20.09	47.04	26.11	41.78
[35]	13.41	43.59	18.67	37.47	21.26	37.35
WCGAN	**11.37**	**40.62**	**15.55**	**33.18**	**19.47**	**33.75**



(a) Input (b) [36] (c) [35] (d) WCGAN

Fig. 11. Qualitative comparison with [35,36].

4.3.3. User study

A user study is performed for a more convincing evaluation. Due to the most recent publication of BBDM, there was insufficient time to incorporate its results into the user study. However, the above qualitative and quantitative analysis have proved WCGAN outperforms BBDM in portrait watercolorization. Fig. 10 presents the statistical results of the comparison with the state of the art methods. In Fig. 10, blue, orange and gray represent three resolutions of 256, 512 and 1024 respectively. The sum of bars with the same color is 100%, where the larger the value, the more competitive this method is at this resolution.

Under three resolutions, we compare WCGAN to four competing methods: Original Gatys, Collection Gatys, Pix2Pix, and I2ICDAE. For each resolution, 10 randomly selected portraits from T_{test} and $T_{multiface}$ are applied to show the performance of all five methods. In each question, we simultaneously display five randomly ordered watercolor pictures generated by different methods, and ask the participants to tick the best watercolor picture based on their subjective feeling. We finally received 1530 votes (51 participants) from two platforms: PC and mobile phone, and more detailed analysis is provided in the supplementary material. The results plotted in Fig. 10 indicate that the WCGAN method, receiving 58.46% votes on average, achieves the best performance among all evaluated methods. Due to their color-mismatching, Original Gatys and Collection Gatys receive the least votes, 6.41% and 8.95% respectively.

4.4. Compared with other multi-scale discriminators

To demonstrate the superiority of the proposed ADA_dis, we conducted comparative experiments by replacing the proposed discriminator architecture in WCGAN with the multi-scale discriminator in [35,36]. Table 5 shows that our proposed ADA_dis quantitatively outperforms the multi-scale discriminators from [35,36] in all evaluation scenarios. The second column of Fig. 11 shows that applying the multi-scale discriminator from [36] results in visible blurring effects due to different branches having the same impact, leading to a compromise among three scales. Applying the multi-scale discriminator from [35] brings about over-stylization in the eye area due to overemphasis on fine-grained features in the latter part of training and reliance on feature feedback by fixed-weight fusion during the whole training. In contrast, our proposed ADA_dis flexibly merges different scale features by activating the strongest responses in different branches through max-pooling.

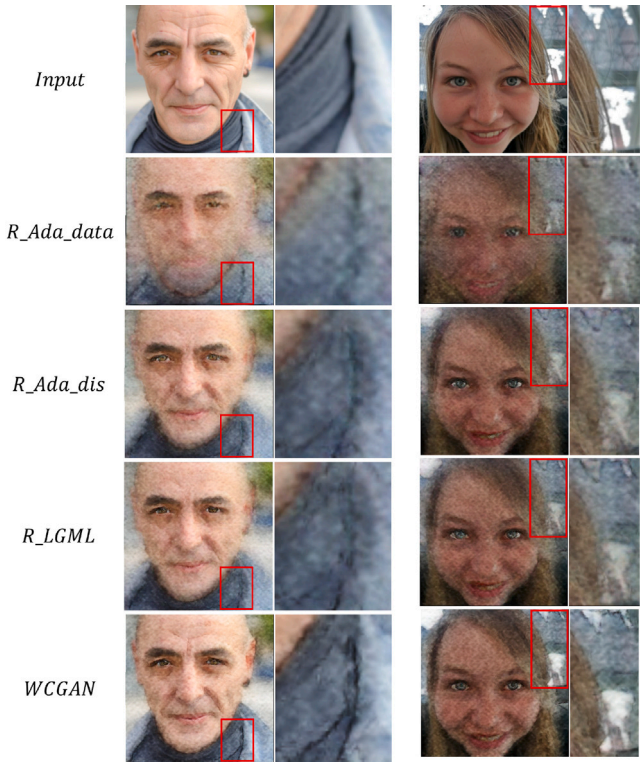


Fig. 12. Ablation studies in Section 4.6.1. Close-ups are provided for the regions in the red boxes.

4.5. Analysis of Local Gram Matrix Loss (LGML)

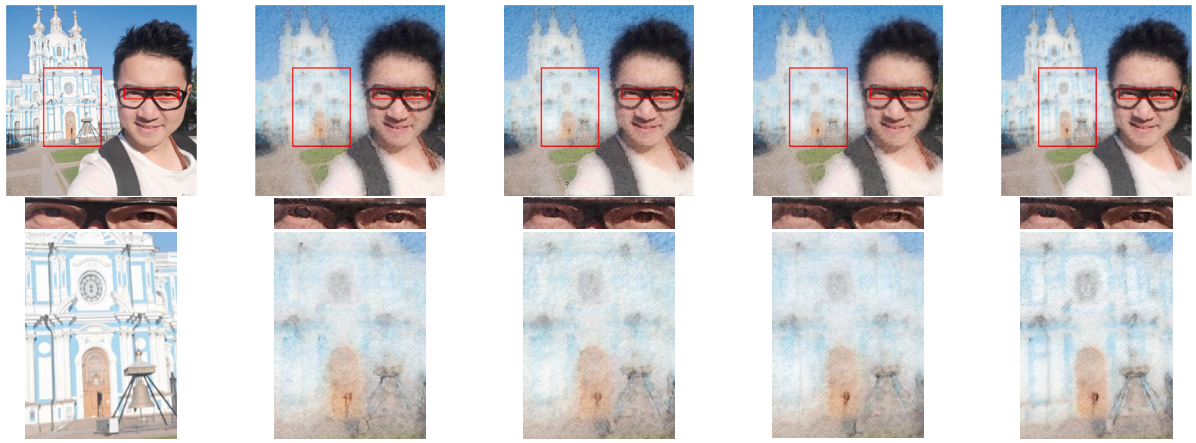
To prove the necessity of the local Gram matrix loss (LGML), we conduct and compare four experiments: R_LGML, *Traditional_GML*, $LGML\_whole$ and WCGAN. WCGAN is the full version of WCGAN; R_LGML removes LGML from WCGAN; *Traditional_GML* replaces LGML applied in WCGAN with traditional Gram matrix loss; $LGML\_whole$ extends the application range of LGML in WCGAN from only the background to the entire picture (including the face). As shown in Fig. 13, compared with the full version WCGAN, R_LGML and *Traditional_GML* lack sufficient capability to fine-tune the stylization quality of local areas, resulting in severe blur in the background. $LGML\_whole$ has a clear drop in the performance of the facial area, especially the eyes.

4.6. Ablation study

The following abbreviations are used to separately represent three ablation study experiments: R_Ada_data for removing $T_{training}$; R_Ada_dis for removing ADA_dis; R_LGML for removing Localized Gram Matrix Loss term. Qualitative analysis is shown in Fig. 12. LPIPS and FID provide quantitative analysis (Table 6).

4.6.1. Qualitative analysis

Fig. 12 separately shows the results of R_Ada_data, R_Ada_dis, R_LGML and WCGAN in the 2nd to 5th rows. For R_Ada_data, the rendering quality of facial parts and detailed areas is significantly reduced. For complicated local areas in the background, R_Ada_dis and R_LGML present a visible drop in the multi-effect fusion and boundary processing compared with WCGAN. The above qualitative analysis demonstrates that each component is essential for the best performance. More examples can be found in the supplementary material.



(a) Content Images (b) *R_GML* (c) *Traditional_GML* (d) *LGML_whole* (e) WCGAN

Fig. 13. Ablation study of L_{LGM} loss. (a) content images, (b) removing Gram matrix loss entirely (R_GML), (c) replacing local Gram matrix loss with traditional (global) Gram matrix loss, (d) expanding local Gram matrix loss to the whole image, (e) our model where local Gram matrix loss is applied to non-face regions.

Table 6
Quantitative analysis of ablation study in Section 4.6.2.

Quantitative analysis of ablation study	LPIPS			FID		
	256	512	1024	256	512	1024
WCGAN	**11.37**	**15.55**	**19.47**	**40.62**	**33.18**	**33.75**
Removing Adaptive Discriminator Architecture	13.50	17.56	20.46	46.93	37.56	36.54
Removing Adaptive Training Dataset	33.68	25.89	20.68	128.84	63.15	36.84
Removing Localized Gram Matrix Loss	12.83	17.34	20.66	44.04	36.62	36.75

4.6.2. Quantitative analysis

Quantitative analysis in the ablation study is conducted through LPIPS and FID. As shown in the first row of Table 6, the full model (WCGAN) containing all techniques reaches the lowest values (i.e., best performance) at any scale, which illustrates the necessity of each component. In the remaining part of this subsection, ablated versions are compared with the full model shown in the first row of the table, and the following conclusions can be made:

- Disabling $T_{training}$ has caused a noticeable quality degradation, especially at 256^2 resolution compared to the full model, an increase of 196.30% under LPIPS and 217.17% under FID.
- R_Ada_dis performs significantly worse than the full model at all comparisons, with an average increase of 12.28% in LPIPS and 12.33% in FID.
- Although its performance outperforms R_Ada_data, R_LGML is worse than the full version WCGAN at any scale on both evaluation criteria.

4.7. Video style transfer

We collected 10 videos from videvo.net as the training data set. Flow2 [43] is applied to calculate the bidirectional optical flow between consecutive video frames. The ground truth with watercolor style are generated by our WCGAN trained by still images. Furthermore, a temporal error measure is defined to reflect the coherence, which is calculated as:

$$E_{\text{temporal}} = \frac{1}{T \cdot I} \sum_{t=1}^{T} \sum_{k=1}^{I} C_k \left(S_t^k - S_t^{dk}\right)^2 \tag{8}$$

where T denotes the total number of consecutive frame pairs. I means the total number of all pixels in the trackable area, which is marked as value 1 in C. C (confidence mask) sets all motion boundaries and occluded regions to value 0 and other regions to value 1. The stylized results S_t and S_{t-1} corresponding to frame F_t at time t and frame F_{t-1} at time $t-1$ are separately generated by G. The desired stylized result S_t^d with temporal consistency at time t is generated by warping S_{t-1} based on pre-calculated optical flow. $E_{temporal}$ takes the average temporal error of all consecutive video frame pairs as the performance of the temporal consistency. Lower $E_{temporal}$ indicates smoother results.

Table 7
Comparison of temporal errors under changing training strategies.

Temporal error	256	512	1024
Without temporal consistency	0.1849	0.2542	0.3771
Real video frames	0.1408	0.2216	0.3602
Multi-crop training data	0.1391	0.2432	0.3554
Both real frames and multi-crop	**0.1339**	**0.2028**	**0.3418**

The temporal errors of four methods under three different scales are shown in Table 7. The first row shows the temporal error of WCGAN aiming for image style transfer, where the biggest errors are achieved under all scales since this version does not take temporal consistency into consideration. WCGAN trained by real video frames or Multi-Crop video training data can both reduce temporal errors as shown in the second and third rows. The versions of WCGAN jointly trained with both real video frames and Multi-Crop video training data achieves the smallest temporal errors under all scales as shown in the fourth row. This demonstrates that real video frames and our created Multi-Crop video training data can both enhance the temporal consistency. Furthermore, these two kinds of video training data provide complementary information and work together effectively to further improve temporal consistency.

5. Conclusion

In this paper, we propose WCGAN, a GAN-based model to transfer portrait images to high-quality watercolor paintings. Local Gram matrix loss enables detailed style characteristics to be properly captured. Moreover, the novel adaptive architecture and adaptive training dataset enable WCGAN to cope with portraits of different sizes. The Multi-Crop video training data further enhances the temporal consistency in video

style transfer tasks. Our experimental results show that WCGAN can faithfully transfer watercolor style to portraits and achieve better temporal performance in video style transfer tasks, outperforming existing state-of-the-art methods. In the future work, facing inconsistent styles in real watercolor datasets, we will explore new methods to deal with mixed watercolor style transfer.

CRediT authorship contribution statement

Hongjin Lyu: Methodology, Software, Validation, Writing – original draft. **Paul L. Rosin:** Methodology, Writing – review & editing, Supervision. **Yu-Kun Lai:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the China Scholarship Council [grant numbers 201806420014]. The study also benefited from Cardiff University's ARCCA computing facilities.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.displa.2023.102530.

References

[1] H. Chen, Y.-Q. Xu, H.-Y. Shum, S.-C. Zhu, N.-N. Zheng, Example-based facial sketch generation with non-parametric sampling, in: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vol. 2, IEEE, 2001, pp. 433–438.

[2] M. Meng, M. Zhao, S.-C. Zhu, Artistic paper-cut of human portraits, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 931–934.

[3] M. Zhao, S.-C. Zhu, Portrait painting using active templates, in: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering, 2011, pp. 117–124.

[4] P.L. Rosin, Y.-K. Lai, Watercolour rendering of portraits, in: Pacific-Rim Symposium on Image and Video Technology, Springer, 2017, pp. 268–282.

[5] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2414–2423.

[6] D. Chen, L. Yuan, J. Liao, N. Yu, G. Hua, StyleBank: An explicit representation for neural image style transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1897–1906.

[7] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[8] H. Wang, X. Liang, H. Zhang, D.-Y. Yeung, E.P. Xing, ZM-Net: Real-time zero-shot image manipulation network, 2017, arXiv preprint arXiv:1703.07255.

[9] H. Zhang, K. Dana, Multi-style generative network for real-time transfer, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 349–365.

[10] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[11] H. Chang, J. Lu, F. Yu, A. Finkelstein, PairedCycleGAN: Asymmetric style transfer for applying and removing makeup, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 40–48.

[12] Y. Chen, Y.-K. Lai, Y.-J. Liu, CartoonGAN: Generative adversarial networks for photo cartoonization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9465–9474.

[13] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, L. Lin, BeautyGAN: Instance-level facial makeup transfer with deep generative adversarial network, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 645–653.

[14] R. Yi, Y.-J. Liu, Y.-K. Lai, P.L. Rosin, APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10743–10752.

[15] M.J. Wilber, C. Fang, H. Jin, A. Hertzmann, J. Collomosse, S. Belongie, BAM! The behance artistic media dataset for recognition beyond photography, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1202–1211.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[17] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

[18] J. Huo, S. Jin, W. Li, J. Wu, Y.-K. Lai, Y. Shi, Y. Gao, Manifold alignment for semantically aligned style transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14861–14869.

[19] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, M.-H. Yang, Universal style transfer via feature transforms, in: Advances in Neural Information Processing Systems, 2017, pp. 386–396.

[20] L. Sheng, Z. Lin, J. Shao, X. Wang, Avatar-Net: Multi-scale zero-shot style transfer by feature decoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8242–8250.

[21] M. Wang, W. Lu, J. Lyu, K. Shi, H. Zhao, Generative image inpainting with enhanced gated convolution and Transformers, Displays 75 (2022) 102321.

[22] X. Li, J. Zhang, Y. Liu, Speech driven facial animation generation based on GAN, Displays 74 (2022) 102260.

[23] S. Shahriar, GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network, Displays (2022) 102237.

[24] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, B. Guo, StyleSwin: Transformer-based GAN for high-resolution image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11304–11314.

[25] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, C. Xu, DF-GAN: A simple and effective baseline for text-to-image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16515–16525.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[27] B. Li, K. Xue, B. Liu, Y.-K. Lai, BBDM: Image-to-image translation with Brownian bridge diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1952–1961.

[28] R. Yi, Y.-J. Liu, Y.-K. Lai, P.L. Rosin, Unpaired portrait drawing generation via asymmetric cycle mapping, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR '20), 2020, pp. 8214–8222.

[29] M. Ruder, A. Dosovitskiy, T. Brox, Artistic style transfer for videos, in: German Conference on Pattern Recognition, Springer, 2016, pp. 26–36.

[30] D. Chen, J. Liao, L. Yuan, N. Yu, G. Hua, Coherent online video style transfer, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1105–1114.

[31] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, W. Liu, Real-time neural style transfer for videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 783–791.

[32] C. Gao, D. Gu, F. Zhang, Y. Yu, ReCoNet: Real-time coherent video style transfer network, in: Asian Conference on Computer Vision, Springer, 2018, pp. 637–653.

[33] B. Amos, B. Ludwiczuk, M. Satyanarayanan, OpenFace: A General-Purpose Face Recognition Library with Mobile Applications, Technical Report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[34] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 2015, arXiv preprint arXiv:1511.07122.

[35] A. Shocher, S. Bagon, P. Isola, M. Irani, InGAN: Capturing and retargeting the "DNA" of a natural image, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4492–4501.

[36] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.

[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[38] X. Wang, Y. Li, H. Zhang, Y. Shan, Towards real-world blind face restoration with generative facial prior, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9168–9178.

[39] Z. Wang, J. Zhang, R. Chen, W. Wang, P. Luo, RestoreFormer: High-quality blind face restoration from undegraded key-value pairs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17512–17521.

[40] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, K.-Y.K. Wong, Progressive semantic-aware style transformation for blind face restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11896–11905.

[41] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, D. Cohen-Or, Encoding in style: a StyleGAN encoder for image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2287–2296.

[42] J. Yoo, H. Eom, Y.S. Choi, Image-to-image translation using a cross-domain auto-encoder and decoder, Appl. Sci. 9 (22) (2019) 4780.

[43] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, FlowNet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.