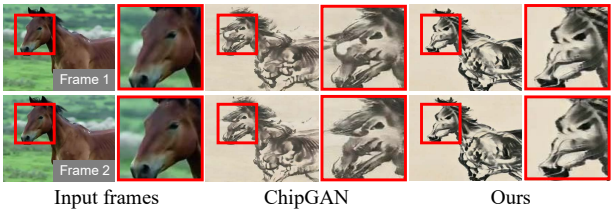
Instance-Aware Coherent Video Style Transfer for Chinese Ink Wash Painting

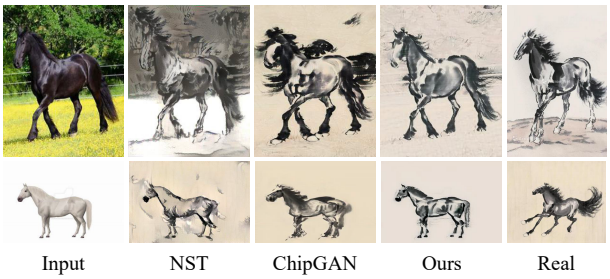
Hao Liang, **Shuai Yang**, **Wenjing Wang** and **Jiaying Liu***
Wangxuan Institute of Computer Technology, Peking University
{lianghao17, williamyang, daooshee, liujiaying}@pku.edu.cn

Abstract

Recent researches have made remarkable achievements in fast video style transfer based on western paintings. However, due to the inherent different drawing techniques and aesthetic expressions of Chinese ink wash painting, existing methods either achieve poor temporal consistency or fail to transfer the key freehand brushstroke characteristics of Chinese ink wash painting. In this paper, we present a novel video style transfer framework for Chinese ink wash paintings. The two key ideas are a multi-frame fusion for temporal coherence and an instance-aware style transfer. The frame reordering and stylization based on reference frame fusion are proposed to improve temporal consistency. Meanwhile, the proposed method is able to adaptively leave the white spaces in the background and to select proper scale to extract feature and depict the foreground subject by leveraging instance segmentation. Experimental results demonstrate the superiority of the proposed method over state-of-the-arts in terms of both temporal coherence and visual quality. Our project website is available at https://obliviousssy.github.io/InkVideo/.



(a) Chinese ink wash video style transfer



(b) Chinese ink wash image style transfer

Figure 1: Our model is able to produce temporally stable Chinese ink wash videos. Moreover, our model can capture the characteristics of Chinese ink wash style better than existing methods.

1 Introduction

Ink video is a modern artistic expression of Chinese ink wash paintings. It inherits the freehand brushwork characteristics of Chinese ink wash paintings, and is widely used in the film, advertising, and art industry. However, to produce traditional ink videos, a large number of paintings need to be drawn, which is considerably time-consuming. This practical requirement motivates our work: we investigate an approach to automatically convert a real video into ink video, which can support large-scale production and help people enjoy, understand and analyze this art form.

Since Gatys *et al.* [2016], there have been many efforts towards image/video stylization. Among them, optical flow is widely used in video stylization for temporal coherence [Ruder *et al.*, 2016; Chen *et al.*, 2017; Huang *et al.*, 2017]. However, existing methods are less effective for ink wash styles. The reasons are twofold: 1) Motions: it is still challenging to estimate optical flow for large motions, while wrong flows will cause flickers and jitters; 2) Strokes: the high contrast between the black strokes and white spaces in ink wash styles makes subtle changes evident, which requires higher temporal consistency. In this paper, we propose a novel multi-frame fusion scheme to solve these problems.

Meanwhile, although methods following [Gatys *et al.*, 2016] show good performance on richly textured styles such as oil paintings, their results are less satisfactory for Chinese ink wash styles as shown in Fig. 1 (b). It is hard for them to learn the ink wash brushstrokes and how to leave white spaces correctly. The challenges of Chinese ink wash style transfer lie in three aspects: 1) White space: Chinese ink wash paintings often leave white spaces to express ethereal beauty, which requires the model to identify the painting subject and determine the areas to leave blank; 2) Contour brushstroke: Chinese ink wash paintings usually do not use filled colors, so the model has to carefully render contour brushstrokes to

*Corresponding author.

distinguish different subjects; 3) Subject size: Subjects in real photos are very diverse in size, which requires the model to adjust the brushstroke size for different subjects, given the fixed kernel size. Recently, attempts have been made for Chinese ink wash style transfer. ChipGAN [He *et al.*, 2018] learns Chinese ink wash style by Generative Adversarial Networks (GANs) [2014]. However, ChipGAN treats the image as a whole and is still not adaptive enough to the semantic contents. In this paper, we propose three specialized instance-aware constraints to meet the above-mentioned challenges.

In general, we propose a novel Chinese ink wash video style transfer network. The key ideas include multi-frame fusion for temporal coherence and instance-aware style transfer. To maintain temporal consistency, we reorder the frames to flexibly select proper reference frames for optical flow estimation, and fuse reference frames at feature-level. Moreover, a spatial transformation consistency constraint is imposed to alleviate the sensitiveness of GANs to small input changes, further stabilizing the outputs. For introducing instance awareness, we first identify the painting subject and background, then apply white space constraint, contour contrast constraint, and adaptive scale selection to guide the network to extract features from the scale adaptive to the image content. In summary, our contributions are threefold:

- We propose a novel instance-aware coherent video style transfer framework, which translates video frames into Chinese ink wash paintings with both high temporal consistency and fine style imitation.
- We propose a multi-frame fusion strategy for robust video style transfer, which is end-to-end trainable and can efficiently generate stable results.
- We analyze the characteristics of Chinese ink wash paintings and propose an instance-aware style transfer method, which can generate proper white spaces and distinct contours, and is adaptive to multi-scale subjects.

2 Related Work

2.1 Image Style Transfer

Style transfer is to transfer the paintings' style to real photos. Traditional methods [Hertzmann *et al.*, 2001] are mainly based on non-parametric texture synthesis, failing to consider the high-level concept of style. Recently, benefiting from the development of deep convolutional networks, Neural Style Transfer (NST) [Gatys *et al.*, 2016] successfully learns high-level style features and shows superior performance. Later, many works improve NST in terms of speed [Johnson *et al.*, 2016; Huang and Belongie, 2017; Li *et al.*, 2017b] and theoretical explanations [Li *et al.*, 2017a]. Besides NST-based methods, another way of style transfer is to treat styles as image domains and use image-to-image translation models such as CycleGAN [Zhu *et al.*, 2017]. However, as analyzed in Sec. 1, the above universal style transfer methods are not suitable for the Chinese ink wash styles, which implies the requirement of specialized architecture and loss designs.

2.2 Video Style Transfer

Directly applying image style transfer to videos leads to flickers and discontinuities. To maintain temporal consistency, Ruder *et al.* [2016] introduce optical flows into NST by adding a new loss function of the continuity of adjacent frames. For acceleration, Chen *et al.* [2017] design a feed-forward video style transfer model based on feature fusion. However, due to large motions, intensive ink wash brushstrokes, and the sensitiveness of GANs, directly exploiting optical flow-based schemes leads to unsatisfactory results. In this paper, we propose a coherent Chinese ink wash video style transfer model based on multi-frame fusion.

2.3 Chinese Ink Wash Painting Generation

Traditional Chinese ink wash style transfer methods [Yu *et al.*, 2003; Yang and Xu, 2013; Liang and Jin, 2013] mainly rely on brushstroke simulation, which is manually designed and not robust enough for real application. ChipGAN [He *et al.*, 2018] is a GAN-based model that learns the ink wash style directly from data. Zhang *et al.* [2020] improve it by modelling styles with AdaIN [Huang and Belongie, 2017]. However, they are less adaptive to the image content, and fail to consider temporal consistency for videos. In this paper, we propose a new model to meet the challenges of Chinese ink wash painting video style transfer.

3 Instance-Aware Coherent Chinese Ink Wash Video Style Transfer

Let X and Y be the real photo domain and Chinese ink wash painting domain, respectively. Our model aims to learn the bidirectional mapping between X and Y, while additionally considering the temporal consistency between consecutive frames. Our model consists of two generators $F : X \rightarrow Y$ and $B : Y \rightarrow X$ with two corresponding discriminators D_Y and D_X. We further disassemble F into two mappings: encoder E and generator G, where E extracts the feature map of the image and G generates ink paintings from the feature map. In this section, we will first introduce our strategy of coherent video stylization. Then, we propose the instance-aware style transfer method for Chinese ink wash paintings. Finally, we show our full training objective.

3.1 Coherent Video Style Transfer

First, we briefly review the idea of optical flow-based temporal consistency constraint. Optical flow describes the motion of each pixel between two frames. Ideally, we have the relationship $x_t = \omega(x_{t-1}, f(x_{t-1}, x_t))$, where x_t is the frame at time t, $f(x, x')$ is the optical flow from frame x to x', and $\omega(x, f)$ represents warping x based on f. An occlusion map M_o is often introduced to solve object occlusion and inaccurate estimation issues: $M_o \otimes x_t = M_o \otimes \omega(x_{t-1}, f(x_{t-1}, x_t))$ with $\otimes$ the element-wise multiplication operator.

There are some drawbacks of the traditional optical flow-based methods. In the following, we will point out and analyze each problem and provide our solutions.

Spatial-Transformation Consistency

Directly warping frames according to optical flow easily causes distortion and seam artifacts [Chen *et al.*, 2017]. Naturally, we warp the feature map instead. However, in experiments, we find that GAN is very sensitive to the feature

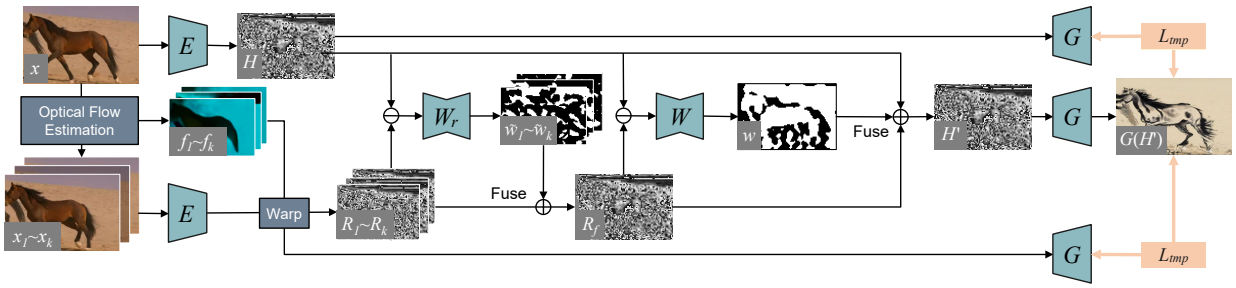


Figure 2: Pipeline of feature-based multi-frame fusion framework. Our method fuses the feature maps from k selected reference frames based on optical flow to achieve temporal coherence. For simplicity, we omit occlusion masks and spatial-transformation consistency loss.

changes, *i.e.*, the transformation on the input feature does not match the transformation on the output image. To solve this problem, we propose a spatial-transformation consistency loss by imposing an input-output consistency constraint under optical flow warping transformation. Let $y_\omega = \omega(F(x), f)$ denote the warped stylized image and $H_\omega = \omega(E(x), f_\downarrow)$ denote the warped feature map, where $f_\downarrow$ is the down-sampled f to match the size of $E(x)$. Our loss function is:

$$L_{cons} = E_x\left[\|(G(H_\omega) - y_\omega) \otimes M_o\|_2^2 / |M_o|\right], \quad (1)$$

where $|\cdot|$ calculates the number of pixels. Intuitively, L_{cons} keeps the consistency of feature input $E(x)$ and image output $F(x)$ under spatial transformations ω.

Feature-Based Multi-Frame Fusion

In most video stylization methods, only the previous frame is used as the reference. However, because of the object occlusion and the inaccurate optical flow estimation, using only one frame may cause large errors. Therefore, we propose a multi-frame feature fusion scheme to stylize the target frame x_t with k reference frames of time $\{t_1, ..., t_k\}$. For simplicity, in the following, we will omit t, and use x and x_i to represent the target frame x_t and the reference frame x_{t_i}, respectively.

As illustrated in Fig. 2, we first align the feature map $H_i = E(x_i)$ of the reference frame x_i to the feature map $H = E(x)$ of the target frame x based on their optical flow: $R_i = \omega(H_i, f_{i\downarrow})$, where $f_{i\downarrow} = f(x_i, x)_\downarrow$ is the down-sampled optical flow between x_i and x. The aligned reference feature maps $\{R_1, ..., R_k\}$ are then fused to form one reference feature map R_f. R_f is further fused with H to obtain the final feature map H', which is decoded to the stylization result $G(H')$. Here, two networks W_r and W are proposed to predict the fusion weight maps. Specifically, W_r is fed with $(R_i - H)$ to predict weight map w_i of R_i. The weight maps are normalized by softmax and used to fuse $\{R_1, \ldots, R_k\}$ to one reference feature map R_f. Similarly, W is fed with $(R_f - H')$ to predict the weight map w, and fuse R_f and H:

$$R_f = \sum\nolimits_{i=1}^{k} \tilde{w}_i \otimes R_i, \text{where } \tilde{w}_i = e^{w_i} / \sum\nolimits_{j=1}^{k} e^{w_j}, \quad (2)$$

$$H' = w \otimes R_f + (\mathbf{1} - w) \otimes H. \quad (3)$$

To train network W_r and W, we propose a temporal loss to require high temporal consistency after fusion. We expect that the stylized frame and the aligned reference frame are consistent in non-occlusion regions:

$$L_{tmp} = E_x\Big[\|(G(H') - F(x)) \otimes M_{o,c}\|_2^2 / |M_{o,c}| + \sum_{i=1}^{k} \|(G(H') - \omega(F(x_i), f_i)) \otimes M_{o,i}\|_2^2 / |M_{o,i}|\Big], \quad (4)$$

where f_i and $M_{o,i}$ are the optical flow and occlusion map between x and x_i. $M_{o,c} = \max(0, 1 - \sum_{i=1}^{k} M_{o,i})$ is the fused occlusion map to indicate regions with no reference.

Frame Selection by Reordering

In our multi-frame fusion process, the k reference frames should contains different information and be fused together. So there is a potential constraint that these reference frames should be "far away" from each other. Meanwhile, sequential reference may accumulate errors in rendering. Errors in the preceding frames will accumulate to subsequent frames through sequential references until the last few results are unacceptable. A non-sequential reference structure such as binary tree may be better. Therefore, we propose a frame reordering scheme. By adjusting the order of generation, we make the reference frames diverse and organize the reference relationships adaptively.

We first divide all frames into two sets: S_{inked} and $S_{uninked}$, denoting frames that have been stylized and to-be stylized, respectively. In each round, we select the most appropriate frame $\hat{x}$ from $S_{uninked}$, and then select k closest frames from S_{inked} as reference frames to generate the corresponding ink video frame of $\hat{x}$. Specifically, to select $\hat{x}$ from $S_{uninked}$, we design the criteria:

$$\hat{x} = \arg\min_x \sum_{x' \in S_{uninked}} \frac{\alpha\delta(x, x')}{|S_{uninked}|} - \sum_{x' \in S_{inked}} \frac{\delta(x, x')}{|S_{inked}|},$$

where α is a hyper parameter, and $\delta(x, x') = \|f(x, x')\|_2^2$ measures the distance between the frames x and x', with the optical flow $f(x, x')$ from x to x'. The whole formula can be understood as that we want to select the frame that is close to the unstylized frames while being different from the stylized frames. So the frames in S_{inked} set are "far away" from each other, which meets our constraint. The proximity of selected frame and unstylized frames means the selected frame will provide reference for more frames in the subsequent rounds.

Then, we select k reference frames from S_{inked}. The selected frames closest to $\hat{x}$ in terms of $\delta(\cdot, \cdot)$ are used as reference. Finally, we generate the ink video frame corresponding

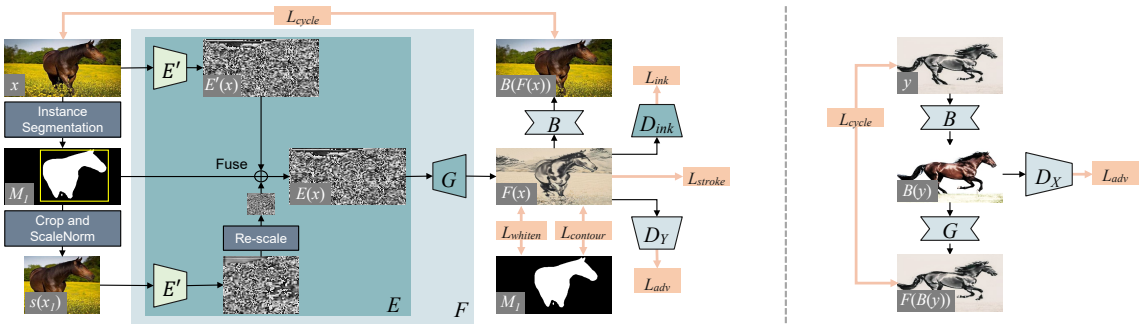


Figure 3: Pipeline of instance-aware Chinese ink wash style transfer. For simplicity, this example only contains one instance ($N = 1$).

to $\hat{x}$ based on the selected reference frames, and then move $\hat{x}$ from $S_{uninked}$ to S_{inked}. Through frame reordering, the reference can be non-sequential, which is better than using sequential reference in terms of avoiding accumulating errors.

3.2 Instance-Aware Chinese Ink Wash Stylization

Chinese ink wash style has many characteristics that need special consideration. By observing professional paintings, we find that the white space is often left out of the painting subject and contour is usually depicted by heavy brushstrokes to highlight the content. Meanwhile, stroke size is basically determined by the size of the subject. For small subjects, finer strokes are used to better depict their details, and vice versa. All these require the identification of the painting subject, which is closely related to object detection.

Therefore, we are motivated to leverage the advanced instance segmentation tool [Kirillov *et al.*, 2019], and propose an instance-aware style transfer network. Based on the detected subjects, we design a whiten loss L_{whiten} for white space and a contour loss $L_{contour}$ for contour brushstroke. Moreover, we propose a scale normalization strategy to adaptively adjust the scale of the subject.

White Space Constraint. Leaving white space means that more ink needs to be allocated to the significant subjects such as the horses in Fig. 1, while leaving trivial objects rough or blank to emphasize the simplicity of the painting. Given the detected foreground instance mask M_f of the target frame $x \in X$, we expect that the ratio of average darkness of background area and foreground area in the generated painting is small. So we propose the whiten loss:

$$L_{whiten} = E_x\left[\frac{\|(\mathbf{1} - F(x)) \otimes M_b\|_2^2 / |M_b|}{\|(\mathbf{1} - F(x)) \otimes M_f\|_2^2 / |M_f|}\right], \quad (5)$$

where $M_b = \mathbf{1} - M_f$ is the background mask. $F(x)$ is the generated painting in RGB space with values in $[-1, 1]$, so $(\mathbf{1} - F(x))$ represents the darkness of the generated painting.

Contour Contrast Constraint. In Chinese ink wash paintings, the brushstrokes need to be carefully distinguished for different objects. To achieve this, we propose to increase the high-frequency components at the contours of the subjects. Specifically, we calculate the contours based on the instance segmentation results, and use a contour loss to increase the differentiation between object boundaries:

$$L_{contour} = 2 - E_x\big[\|Haar(F(x)) \otimes M_c\|_1 / |M_c|\big], \quad (6)$$

where M_c is the contour mask of x, with 1 represents the subject contour and 0 represents others. $Haar(x)$ is the high-frequency components of x by applying the Haar wavelet, with values in $[-2, 2]$. Thus, offset 2 is added in the loss.

Adaptive Scale Selection. Since deep models are limited by the convolution kernel size, it is difficult to adaptively extract features according to the subject scale. Therefore, we design a novel scale normalization strategy. Considering frame x with N instances, let x^n and M^n be the bounding box region and the mask of the n-th subject, respectively, *i.e.*, $M_f = \sum_{n=1}^{N} M^n$. As shown in Fig. 3, our encoder E consists of a basic encoder E'. We first scale each subject region to a predefined fixed size, obtaining $s(x^n)$ with s the scale operation. Then we feed it to E' to extract the feature map under this predefined scale. Later, we re-scale the feature map H_b^n back to its original size as $H^n = s^{-1}(H_b^n)$ with s^{-1} the re-scale operation, and fuse all H^n with the original feature map $E'(x)$ to obtain the final feature map $E(x)$:

$$E(x) = \sum_{n=1}^{N} H^n \otimes M^n + E'(x) \otimes M_b. \quad (7)$$

The final stylization result is obtained as $F(x) = G(E(x))$. Here, the fixed size is set to the averaged subject size in the training set, which means the normalized subject is stylized using the most suitable brushstroke size. In Sec. 3.1, we introduce a spatial-transformation consistency constraint. Now we can extend it to both warping and scale transformations:

$$\begin{aligned} L_{cons} &= E_x\big[\|(G(H_\omega) - y_\omega) \otimes M_o\|_2^2 / |M_o|\big] \\ &+ E_x\Big[\sum_{n=1}^{N} \|(s^{-1}(G(H_b^n)) - F(x)) \otimes M^n\|_2^2 / |M^n|\Big]. \end{aligned} \quad (8)$$

3.3 Full Objective

We use the adversarial cycle loss [Zhu *et al.*, 2017] to learn the bidirectional mapping between X and Y:

$$\begin{aligned} L_{adv} &= E_y\big[\log D_Y(y)\big] + E_x\big[\log(1 - D_Y(F(x)))\big] \\ &+ E_x\big[\log D_X(x)\big] + E_y\big[\log(1 - D_X(B(y)))\big], \\ L_{cycle} &= \gamma E_x\big[\|B(F(x)) - x\|_1\big] + E_y\big[\|F(B(y)) - y\|_1\big]. \end{aligned}$$

Converting from photos to paintings would inevitably lose some information. Therefore, we add a parameter $\gamma < 1$ to reduce the forward cycle consistency in L_{cycle}.

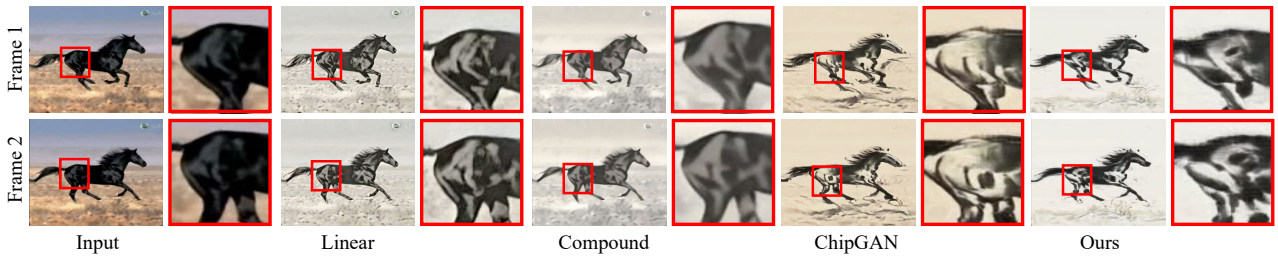


Figure 4: Video style transfer comparison with ChipGAN [2018], Linear [2019] and Compound [2020].

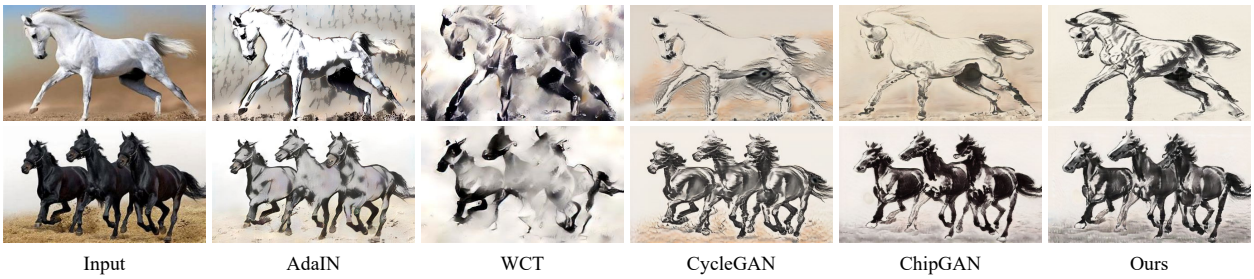


Figure 5: Comparison with AdaIN [2017], WCT [2017b], CycleGAN [2017] and ChipGAN [2018] on single frame stylization.

Method	Linear	Compound	ChipGAN	Ours
Temporal	2.95	**3.13**	1.73	2.19
Stylization	1.72	1.49	3.16	**3.63**
Comprehensive	1.83	1.65	3.02	**3.50**

(a) Chinese ink wash video style transfer

Method	AdaIN	WCT	CycleGAN	ChipGAN	Ours
MOS	2.72	2.00	2.73	3.14	**4.40**

(b) Chinese ink wash image style transfer

Table 1: User preference rate of different methods.

We also incorporate the brushstroke loss L_{stroke} and ink wash loss L_{ink} introduced by ChipGAN [He *et al.*, 2018] to emphasize the subject edges and realize diffusion effects of the ink wash, respectively. Finally, our full objective is:

$$\begin{aligned} L =& L_{adv} + \lambda_1 L_{cycle} + \lambda_2 L_{cons} + \lambda_3 L_{tmp} \\ &+ \lambda_4 L_{whiten} + \lambda_5 L_{contour} + \lambda_6 L_{stroke} + \lambda_7 L_{ink}, \end{aligned} \tag{9}$$

where λ_i are hyper-parameters to balance these losses. We will later verify their effects through ablation studies.

4 Experimental Results

4.1 Implementation Details

For training, we use Chinese ink wash paintings in ChipPhi dataset [He *et al.*, 2018], and further collect 115 videos of horses from the Internet, about 10k frames in total. PWC-Net [Niklaus, 2019; Sun *et al.*, 2018] is used to estimate optical flows for training and testing. More experimental details and results are provided in the supplementary material.

4.2 Comparison with State-of-the-Arts

To validate the effectiveness of the proposed method, we compare with the following state-of-the-art video style transfer methods: Linear [Li *et al.*, 2019], Compound [Wang *et al.*, 2020], and image style transfer methods: AdaIN [Huang and Belongie, 2017], WCT [Li *et al.*, 2017b], CycleGAN [Zhu *et al.*, 2017], ChipGAN [He *et al.*, 2018].

Method	AdaIN	WCT	Linear	CycleGAN	ChipGAN	Ours
FID	250.98	302.04	250.73	262.6	216.54	**213.16**

Table 2: FID score of different methods.

Video Style Transfer. Results are shown in Fig. 1(a) and Fig. 4. ChipGAN does not consider temporal consistency, therefore the results have flickers and jitters. Linear and Compound avoid the problem of flickering. However, they fail to imitate the representative ink strokes, and suffer from color bias and blurry artifacts. By comparison, our model can maintain good temporal consistency thanks to the proposed schemes. Videos and more results can be found in the supplementary material. For quantitative comparison, we invite 40 users to score the results from three aspects of temporal consistency, stylization quality, and comprehensive quality. In Table 1(a), our method has the highest overall quality and stylization score.

Image Style Transfer. Results are shown in Fig. 1(b) and Fig. 5. AdaIN and WCT transfer the color and tonality but fail to render brushstrokes. CycleGAN tends to use dry strokes to depict the image content. The brushstrokes are messy and lack variety. ChipGAN fuses the foreground subjects and surrounding backgrounds, with strokes interwoven. By comparison, our model successfully leaves white spaces in the background and better depicts the contour of each instance subject. Our results are clean and of high contrast, following the simplicity principle of Chinese ink wash paintings. For quantitative evaluation, Table 1(b) reports Mean Opinion

Config	w/o fusion	w/o reordering	full
Temporal	1.92	1.72	**2.36**
Stylization	1.91	1.92	**2.18**
Comprehensive	1.88	1.81	**2.32**

(a) Chinese ink wash video style transfer

Config	w/o L_{whiten}	w/o $L_{contour}$	w/o scaleNorm	full
MOS	1.83	2.21	2.95	**3.01**

(b) Chinese ink wash image style transfer

Table 3: User preference rate for ablation studies.

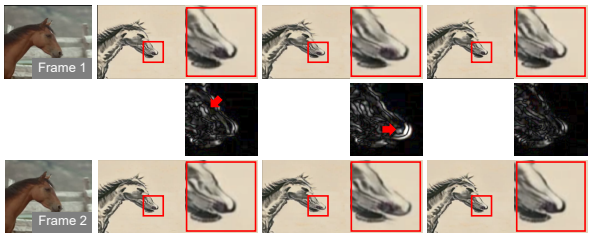


Figure 6: Effects of frame fusion and frame reordering. The difference between the two frames is visualized in the middle row.

Score (MOS) in the user study and Table 2 reports the FID score [Heusel *et al.*, 2017], where our model is the best.

4.3 Ablation Study

In Fig. 6, we compare the video style transfer results under different configurations. Without frame fusion, the result has texture jitters. Without frame reordering, the horse's nose becomes longer as the horse moves , which is caused by sequential references. By comparison, our full model avoids these problems and achieves better temporal consistency.

Fig. 7 analyzes the designs of our instance-aware Chinese ink wash stylization. Without L_{whiten}, the result is much messy. Without $L_{contour}$, the contour of the horse gets blurred and mixed with the background. Without scale normalization, the model fills the body of the horse with black, which is rare in Chinese ink wash paintings. In comparison, the result of our full model is more vivid and balancing.

For quantitative comparison, we conduct a user study in Table 3 following the previous settings. The full model has the best result, demonstrating the effectiveness of our designs.

4.4 Model Generalization

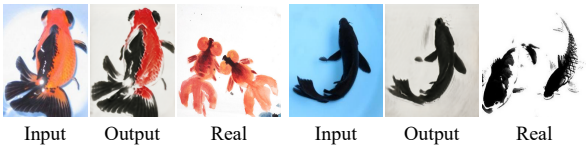
Generalization of Content. Though trained only on horses, our model has learned the general feature of ink wash, thus can transfer other subjects as shown in Fig. 8(a). We also collect data and train for fish. The results are good as shown in Fig. 8(b). More results and details are in supplementary.

Generalization of Style. The proposed model is not limited to Chinese ink wash paintings. It can be well applied to other styles. In Fig. 8(c), we trained our model on Quick Draw Dataset[1]. Our model successfully generates sketches better

[1]https://github.com/googlecreativelab/quickdraw-dataset

Figure 7: Ablation studies for instance-aware style transfer.

(a) Results on dog and tiger with model trained on horses.



(b) Results of training on fish dataset

(c) Sketch painting style transfer.

Figure 8: Generalization of the proposed method.

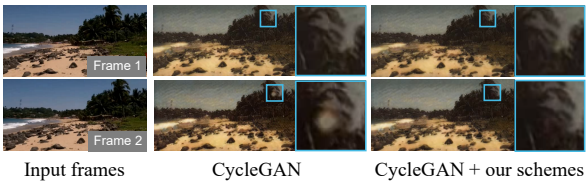


Figure 9: The improvement brought by our method to CycleGAN.

than CycleGAN. Moreover, in Fig 9, our frame fusion and reordering schemes effectively improve the temporal consistency of CycleGAN for oil painting video style transfer.

5 Conclusion

In this paper, we study a challenging problem of video style transfer for Chinese ink wash painting. A novel multi-frame fusion with frame reordering is proposed for temporal consistency, and an instance-aware style transfer is proposed to transfer ink wash painting characteristics. We validate the effectiveness and robustness of our methods by comparison experiments and comprehensive ablation studies.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China under Contract No.61772043 and a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[Chen *et al.*, 2017] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proc. Int'l Conf. Computer Vision*, pages 1105–1114, 2017.

[Gatys *et al.*, 2016] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[He *et al.*, 2018] Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *Proc. ACM Int'l Conf. Multimedia*, pages 1172–1180, 2018.

[Hertzmann *et al.*, 2001] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Proc. the 28th Annual Conf. Computer Graphics and Interactive Techniques*, pages 327–340, 2001.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626—-6637, 2017.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int'l Conf. Computer Vision*, pages 1510–1519, 2017.

[Huang *et al.*, 2017] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 783–791, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Fei Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conf. Computer Vision*, pages 694–711, 2016.

[Kirillov *et al.*, 2019] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. 2019.

[Li *et al.*, 2017a] Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Int'l Joint Conf. Artificial Intelligence*, pages 2230–2236, 2017.

[Li *et al.*, 2017b] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 386–396, 2017.

[Li *et al.*, 2019] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2019.

[Liang and Jin, 2013] Lingyu Liang and Lianwen Jin. Image-based rendering for ink painting. In *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics*, pages 3950–3954, 2013.

[Niklaus, 2019] Simon Niklaus. A reimplementation of PWC-Net using PyTorch. https://github.com/sniklaus/pytorch-pwc, Nov 19, 2019.

[Ruder *et al.*, 2016] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German Conf. Pattern Recognition*, pages 26–36, 2016.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2018.

[Wang *et al.*, 2020] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *AAAI Conference on Artificial Intelligence*, 2020.

[Yang and Xu, 2013] Lijie Yang and Tianchen Xu. Animating chinese ink painting through generating reproducible brush strokes. *Science China Information Sciences*, 56(1):1–13, 2013.

[Yu *et al.*, 2003] Jinhui Yu, Guoming Luo, and Qunsheng Peng. Image-based synthesis of chinese landscape painting. *Journal of Computer Science and Technology*, 18(1):22–28, 2003.

[Zhang *et al.*, 2020] Fengquan Zhang, Huaming Gao, and Yuping Lai. Detail-preserving cyclegan-adain framework for image-to-ink painting translation. *IEEE Access*, 8:132002–132011, 2020.

[Zhu *et al.*, 2017] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. Int'l Conf. Computer Vision*, pages 2242–2251, 2017.