Local Consistency Guidance: Personalized Stylization Method of Face Video (Student Abstract)

Wancheng Feng[1] , **Yingchao Liu**[1] , **Jiaming Pei**[2] , **Wenxuan Liu**[1] , **Chunpeng Tian**[1] , **Lukun Wang**[1*]

[1]Shandong University of Science and Technology, Taian Shandong China
[2]University of Sydney, Sydney Australia
cnfengwancheng@gmail.com, jpei0906@uni.sydney.edu.au,
liuyingchao,liuwenxuan,tianchunpeng,wanglukun@sdust.edu.cn

Abstract

Face video stylization aims to convert real face videos into specified reference styles. While one-shot methods perform well in single-image stylization, ensuring continuity between frames and retaining the original facial expressions present challenges in video stylization. To address these issues, our approach employs a personalized diffusion model with pixel-level control. We propose the Local Consistency Guidance (LCG) strategy, composed of local cross-frame attention and inter-framed style transfer, to ensure temporal consistency. This framework enables the synthesis of high-quality stylized face videos with excellent temporal continuity.

Introduction

Face video stylization has made significant strides in recent years, driven primarily by the adoption of GAN-based approach (Chong and Forsyth 2022). However, achieving high-quality face video stylization remains a challenge, with our paper addressing three key issues: Firstly, existing one-shot methods struggle to capture the richness of facial expressions in videos due to the limited diversity of stylized samples. Secondly, GAN-based methods require face alignment before performing 1:1 stylization, yet facial representations in videos are dynamic. Finally, the introduction of DreamBooth (Ruiz et al. 2023) opens up possibilities for personalized generation. However, ensuring both style-level and detail-level consistency between frames remains a challenge.

In addressing these challenges, we propose a novel face video stylization framework incorporating Local Cross-Guidance (LCG). LCG helps handle inter-frame discontinuities and enhances expression richness during generation. We introduce a local cross-attention mechanism and local style transfer for optimizing local information, achieving high-quality video stylization with good continuity through iterative iterations.

In the inference stage of our UNet model, we employ local cross-attention strategies to guide information exchange between frames, enhancing style-level consistency in a naturally continuous frame generation strategy.

*Corresponding author.

For denoising, we introduce a local style transfer strategy to maintain detail-level consistency. We use a traditional example-based style transfer method, effectively preserving detail-level consistency. However, for frames with significant variations, we draw inspiration from its successful application to local frames, incorporating it into iterative denoising inference to reduce potential noise.

Methods

The Personalized Diffusion Model is a diffusion-based method with Local Cross-Guidance (LCG) constraints, comprising primarily two components: local-cross attention and local style transfer. Overwork can be seen in 1 (a).

For local cross attention. To ensure the accuracy of local information and inspired by other zero-shot video editing techniques (Ceylan, Huang, and Mitra 2023), this paper replaces the original self-attention layers in U-Net with local-cross attention. Specifically, we provide a window that multiplies the local Q, K, V by a weight decreasing at the center node, and the center node moves with the window. Between adjacent frames in this way, can ensure that the generated style is more close to, so as to promote the continuity of video.

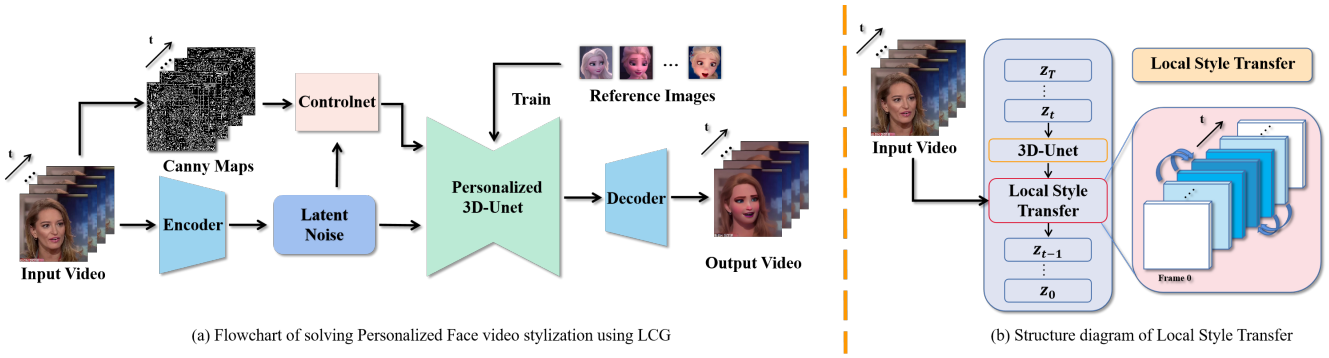
For local style transfer. Utilizing local-cross attention enhances personalized style control for improved video continuity (Jamriska 2018). However, it may not fully achieve seamless transitions between frames. We've observed that example-based style transfer works well for preserving detail consistency but may accumulate deviations in frames with larger temporal gaps. Nevertheless, it excels at the local level. Building on this, we propose extending its application from local to global contexts. See 1 (b) for the structure of local style transfer.

We define the original input as x_t^f, style images are the images at denoising step t, defined as s_t^f, and the images to be stylized as x_t^{f+1}. In other words, the transfer process from frame f to f+1 can be described as:

$$I_f^{f+1} = Transfer(x_t^f; x_t^{f+1}, s_t^f) \tag{1}$$

Where $Transfer(\cdot)$ means the example-based style transfer framework.

We first determine the number of neighboring nodes. Then, each neighboring node's function is weighted expo-



(a) Flowchart of solving Personalized Face video stylization using LCG

(b) Structure diagram of Local Style Transfer

Figure 1: Overall of our proposed Framework. (a) The video is first used to form latent noise by encoder, and then combined with controlnet and trained Dreambooth as input to the Personalized Diffusion Model (Local Consistency is carried out in this part Guidance) and finally get the stylized video. (b) The structure of Inter-framed Style Transfer, which is a process to enforce the detail-level consistency on the video when DDIM Denoising.

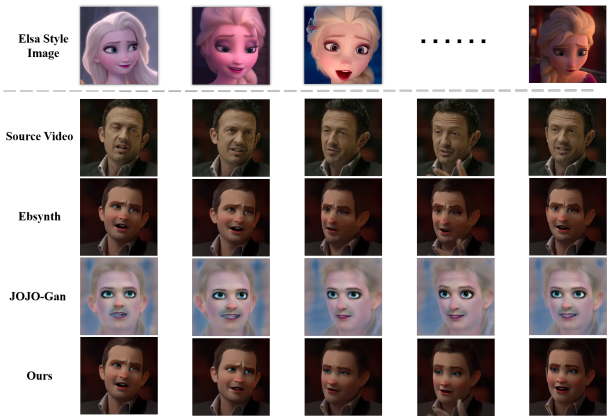


Figure 2: Our method compares the visualization results of Ebsynth as well as JOJOGan method.

nentially based on its distance from the center node using the mentioned transfer method. Finally, the original latent content is replaced with these weighted values.

Experiments

In this section, we present experimental results using the public dataset VFHQ (Xie et al. 2022) for testing purposes. We utilized Ebsynth for example-based style transfer and JOJOGan for compare experiments. To ensure fairness, we employed the first frame generated by our method as the stylized sample image when using Ebsynth. As shown in Figure 2, our approach seamlessly produces results that accurately capture distinctive facial features and makeup of the test sample, while also providing control over expressions. Our method, through the Local Zero-Shot strategy, effectively addresses limitations found in existing one-shot methods, particularly in video applications, demonstrating its overall effectiveness. At the same time, we also conducted a quantitative analysis of the generated images with other works, and we mainly compared the value of frame consistency, as detailed in 1.

Method	Frame Consistency(%)
Source video	0.91
Dreambooth+Controlnet	0.83
Ours	0.97

Table 1: Comparison of our work with other generation method in computing Frame Consistency.

Conclusion

We introduce Local Consistency Guidance (LCG), a novel approach to the face video stylization problem, presenting a comprehensive framework for face stylization. We validate the feasibility of our work through result visualization. In the future, we plan to extend our research to encompass a broader range of character styles and further optimize and fine-tune parameters to enhance content diversity.

References

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*.

Chong, M. J.; and Forsyth, D. 2022. Jojogan: One shot face stylization. In *European Conference on Computer Vision*, 128–152. Springer.

Jamriska, O. 2018. Ebsynth: Fast Example-based Image Synthesis and Style Transfer. https://github.com/jamriska/ebsynth. Accessed: 2019-May-10th.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Xie, L.; Wang, X.; Zhang, H.; Dong, C.; and Shan, Y. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 657–666.