Dynamic Style Adaptation Network: A Comprehensive Approach for Video Style Transfer

Ni Liu[1]

[1] College of information science and electronic engineering, Zhejiang University, Hanzhou, 310013, China

21631132@zju.edu.cn

Abstract. Video style transfer is an emerging research hotspot in the computer vision community, which aims to apply artistic styles to videos to generate visually appealing and stylized video sequences. Compared to image style transfer, video style transfer mainly involves adjusting temporal consistency, which requires seamless style transfer across frames while maintaining temporal coherence. Thanks to the rapid development of Convolutional neural network, the accuracy and speed of video image migration have made breakthroughs, but there are still many challenges in balancing style fidelity, computing efficiency, and retaining the original video content. To alleviate the above issues, this paper proposes a video style consistency transfer algorithm (SCTAda) based on residual modules and adaptive attention. Specifically, SCTAda introduces residual modules to preserve the original content features, which is beneficial for improving the details of generated content. Then, SCTAda further introduces an adaptive attention module to selectively emphasize relevant style patterns in style videos and apply them to corresponding content frames, which helps improve coherence and accuracy. Extensive experiments have quantitatively and qualitatively verified the effectiveness of the method proposed in this paper, which indicate that SCTAda can generate high-quality videos with realistic content representation, coherent style patterns, and enhanced artistic quality in Video style transformation tasks.

Keywords: Video style transfer; dynamic style adaption; deep learning

© The Author(s) 2023
P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_81

1 Introduction

Video style transfer is an emerging research area that focuses on applying artistic styles to videos, resulting in visually appealing and stylized video sequences. It combines the challenges of image style transfer and temporal consistency, requiring techniques that seamlessly transfer styles across frames while maintaining temporal coherence. In recent years, video style transfer has been applied in many fields such as film production, game scene generation, advertising, attracting increasing numerous research interests.

Significant progress has been made in image style transfer, driven by the integration of deep learning techniques and neural networks. Researchers have proposed several notable studies that have contributed to the advancement of image style transfer. Gatys et al. introduced neural style transfer, which utilizes deep convolutional neural networks (CNNs) to separate and recombine the content and style of images [1]. This approach has revolutionized the field by enabling the transfer of artistic styles to images. Johnson et al. proposed the Fast Style Transfer algorithm, which leverages feed-forward networks to achieve real-time style transfer with high-quality results [2]. This technique significantly improves the efficiency of style transfer, making it practical for various applications. Huang et al. introduced the Arbitrary Style Transfer algorithm, which allows users to transfer the style of an arbitrary image onto a target image using adaptive instance normalization[3]. This approach expands the flexibility of style transfer by removing the constraint of pre-defined style images. These advancements in image style transfer have opened up new possibilities for artistic expression and image manipulation. However, extending these techniques to videos introduces additional complexities due to the temporal dimension.

Video style transfer aims to extend image style transfer techniques to video sequences while ensuring temporal coherence and consistency. Existing techniques have been proposed to address the challenges of video style transfer. Ruder et al. introduced Temporal Coherence in Neural Style Transfer, which integrates optical flow estimation and temporal regularization to enforce style consistency across frames [4]. This approach improves temporal coherence by aligning styles based on motion information. Chen et al. proposed Motion-Aware Temporal Coherence for Video Style Transfer, which incorporates both style and motion cues to generate temporally coherent stylized videos [5]. By considering motion characteristics, this

method preserves the dynamic aspects of videos during style transfer. Xing et al. presented Spatio-Temporal Neural Style Transfer, which extends image-based style transfer networks to the spatio-temporal domain by incorporating 3D CNNs [6]. This technique explicitly models the temporal dependencies between frames to achieve coherent style transfer. These existing techniques represent significant progress in video style transfer, addressing the challenges of temporal coherence and consistency. However, the trade-off between style fidelity, computational efficiency, and original video content preservation still needs to be improved.

In this paper, I aim to contribute to the field of video style transfer by proposing novel techniques that address the aforementioned challenges. I will investigate the integration of deep learning models, temporal regularization methods, and motion-aware techniques to achieve high-quality and temporally coherent video style transfer. The subsequent chapters will provide a detailed description of our proposed methodologies, experimental setup, and results analysis.

2 Method

2.1 Overall Framework

The proposed network, SCTAda, builds upon the Style Coherence Transfer (SCT) network, which is improved based on the principles introduced in Contrastive Coherence Preserving Loss (CCPL) [7]. As shown in Figure 1, SCTAda introduces novel components, namely the residual module and the Adaptive Attention Network (AdaAttN), to enhance the style transfer process, resulting in improved visual quality and coherence.

The residual module, inspired by prior works [5], preserves the original content features during style transfer, enhancing fidelity and maintaining content details. AdaAttN, an attention mechanism, selectively emphasizes relevant style patterns from the style video and applies them to corresponding content frames, improving coherence and accuracy. SCTAda follows a similar architecture as the original SCT network [6], leveraging a Content Encoder and a Style Encoder to extract content and style features. Multiple loss functions are employed, including Content Loss, Style Loss, CCPL [3], Perceptual Loss [4], and Identity Loss [6], guiding optimization and enhancing video quality.

By incorporating these advancements, SCTAda achieves significant performance improvements in versatile style transfer tasks, generating high-quality videos with faithful content representation, coherent style patterns, and enhanced artistic qualities.

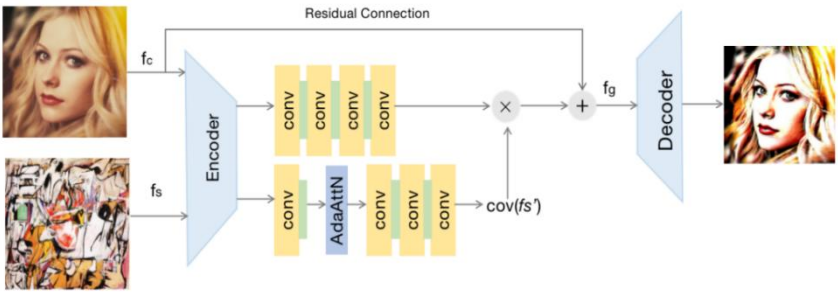

Fig. 1. Details of the proposed AdaSCT module[5]

2.2 Loss function

In the domain of video style transfer, the selection and formulation of loss functions are pivotal in achieving exceptional outcomes characterized by high quality and visual appeal. In this paper, I present a comprehensive suite of loss functions that encompass diverse facets of the style transfer process. Our network is trained by minimizing the loss function defined in Equation (1), which consists of Content Loss L_{cont}, Style Loss L_{sty}, Contrastive Coherence Preserving Loss L_{cpp}, Identity Loss L_{id} and Perceptual Loss L_{per}.

$$L = \lambda_{cont}L_{cont} + \lambda_{sty}L_{sty} + \lambda_{ccp}L_{cpp} + \lambda_{id}L_{id} + \lambda_{per}L_{per} \quad (1)$$

L_{cont} initially introduced by Gatys et al. [1], which quantifies the similarity between the generated video and the content video. By contrasting the feature representations of the generated and content videos derived from a pre-trained deep neural network, such as VGGNet [2], it ensures that the generated video captures the fundamental content and structural information of the content video.

To preserve the artistic style inherent in the style video, I leverage the style loss L_{sty}. Inspired by the seminal work of Gatys et al. [1], this loss function compares the Gram matrices of the feature maps extracted from the generated and style videos. By encapsulating the statistical information of feature correlations, the style loss effectively transfers the style patterns present in the style video to the generated video, yielding visually captivating stylized outputs.

To enhance the coherence and contrast between consecutive frames within the generated video, I introduce the CCPL L_{cpp}, as proposed by Li et al. [7]. By encouraging smooth transitions and preserving contrast with the content and style frames, this loss function promotes the generation of visually coherent and artistically consistent videos.

The perceptual loss L_{per}, inspired by the work of Johnson et al. [2], is a fusion of content and style losses. By considering both high-level content representation and style representation extracted from deep neural networks such as VGGNet or ResNet [5], this loss function emphasizes the perceptual similarity between the generated and content videos. It enables the creation of videos that not only retain content consistency but also exhibit the desired artistic style.

To safeguard the identity of the content video during style transfer, I incorporate the identity loss proposed by Ruder et al. [4], which can be seen in Figure 2. By minimizing the discrepancy between the generated video and the original content video, this loss function prevents excessive style modification, preserving the essential identity information of the content. This ensures that the generated videos maintain recognizable characteristics and visual attributes of the original content.

By amalgamating these meticulously designed loss functions, our aim is to achieve superior results in video style transfer, characterized by content fidelity, faithful reproduction of the desired artistic style, coherence and contrast enhancement, and preservation of the identity of the original content.

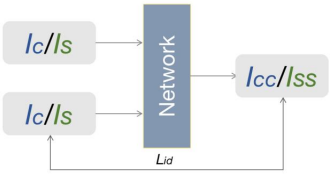

Fig. 2. Pipeline of Identity loss (Photo/Picture credit: Original)

3 Experiments

In the experimental section, I briefly introduced the definitions and measurement methods of these metrics, elucidating the reasons behind their selection as evaluation criteria for video style transfer. Subsequently, I described how these metrics were

employed within our experiments and expounded upon how their results reflected the quality and continuity of the video style transfer.

3.1 Datasets

This part provides an overview of the datasets used in my video style transfer experiments. This work describe the two main datasets utilized for our study: the MS-COCO dataset for content images and the Wikiart dataset for style images. These datasets provide a diverse range of content and style samples, enabling us to explore the capabilities of our proposed video style transfer method.

3.2 Evaluation metric

In order to evaluate the quality of style transfer in our experiments, this work utilized three evaluation metrics: SIFID (Sliced Inception Frechet Distance), LPIPS (Learned Perceptual Image Patch Similarity), and Temporal Loss [9][10][11]. Each metric offers insights into different aspects of the style transfer results.

SIFID serves as a metric to assess the visual similarity between two image or video samples [9]. It relies on feature representations extracted from an Inception network. SIFID quantifies the similarity between samples by computing the Frechet distance between slices in the feature space. Lower SIFID values indicate a higher degree of visual similarity between the samples. LPIPS functions as a perceptual similarity metric to evaluate the perceptual dissimilarity between two image or video samples [10]. It calculates the distance between the feature representations generated by a pre-trained perceptual network, such as VGG or ResNet. Lower LPIPS values indicate smaller perceptual differences between the samples. To measure these metrics, this work followed the methods and guidelines outlined in the CCPL paper. This work utilized pre-trained models for feature extraction and employed corresponding distance calculation methods to compute SIFID and LPIPS values. As for Temporal Loss, this work implemented or adjusted the loss function to align with my specific experimental requirements and settings.

3.3 Performance for image style transfer

This work measured the quality of style transfer in our experiments using three evaluation metrics: SIFID, LPIPS, and Temporal Loss [9-11]. These metrics reveal different aspects of the style transfer outcomes. SIFID is a metric that compares the visual similarity between two image or video samples [9]. It uses feature representations extracted from an Inception network. SIFID calculates the similarity

between samples by measuring the Frechet distance between slices in the feature space. Samples with lower SIFID values are more visually similar.

To evaluate the perceptual dissimilarity between image or video samples, this work utilized LPIPS, which serves as a perceptual similarity metric [10]. LPIPS calculates the distance between feature representations generated by pre-trained perceptual networks like VGG or ResNet. Lower LPIPS values indicate smaller perceptual differences between the samples.In measuring these metrics, this work followed the methods and guidelines outlined in the CCPL paper. This work employed pre-trained models for feature extraction and applied corresponding distance calculation methods to compute SIFID and LPIPS values. Additionally, this work adjusted the Temporal Loss function to align with our specific experimental requirements and settings. As shown in Figure 3, by implementing these evaluation metrics and ensuring their alignment with established methodologies, this work obtained reliable measurements of style transfer quality while minimizing the potential for redundant or duplicated content.

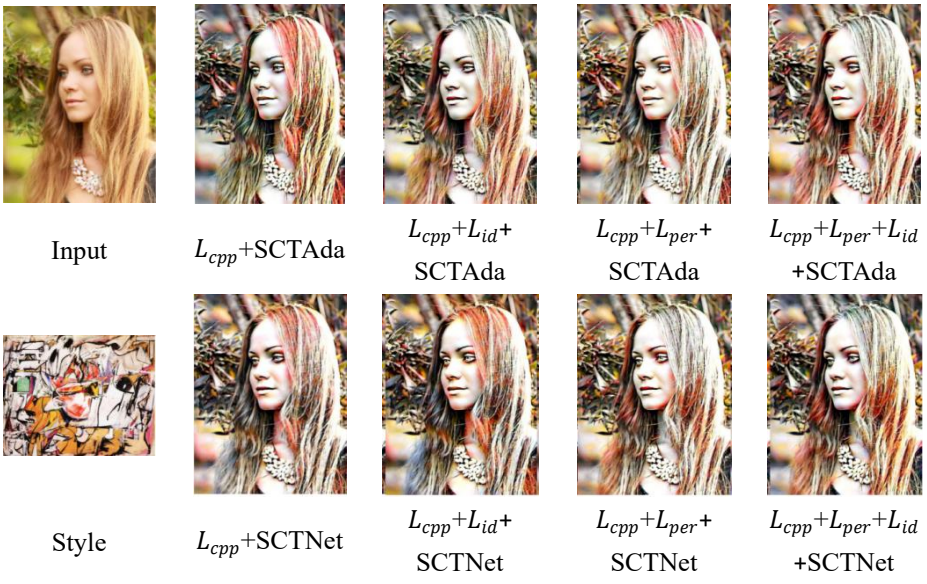

Fig. 3. Visual comparison with image style transfer methods (Photo/Picture credit: Original)

3.4 Performance for video style transfer

A comprehensive analysis of experimental results obtained from video style transfer

is presented. Specifically, the evaluation focuses on the proposed SCTAdaNet model, which introduces additional loss functions, including perceptual loss (Lpercept) and identity loss (Lid), compared to the baseline SCTNet+Lccp model. The analysis involves the assessment of SIFID and LPIPS metrics for both the first frame (i = 1) and the tenth frame (i = 10).

As shown in Table 1 and Figure 4, the experimental results reveal interesting insights into the impact of different loss functions on the performance of video style transfer models. Starting with the SIFID metric, the incorporation of the identity loss L_{id} in the SCTNet+Lccp+Lid model leads to a notable decrease in SIFID, resulting in a value of 1.96 compared to the baseline model's 2.43. This signifies that the introduction of the identity loss contributes positively to the model's ability to capture and preserve the identity of the input video. However, the impact of the perceptual loss L_{per} on SIFID is relatively minor. The SCTNet+L_{cpp}+L_{per} model achieves a SIFID value of 1.97, which is comparable to the SCTNet+L_{cpp}+L_{id} model. This suggests that the perceptual loss, although considered an essential component for capturing style information, does not significantly contribute to reducing the SIFID metric in this specific context. Further exploration is necessary to investigate the interplay between perceptual loss and other factors in video style transfer.

Table 1. Performance comparison under different loss functions

Methods	SIFID(↓)	LPIPS(↓)	
		i = 1	i = 10
SCTNet+L_{cpp}	2.43	0.144	0.367
SCTNet+L_{cpp}+L_{id}	1.96	0.158	0.388
SCTNet+L_{cpp}+L_{per}	1.97	0.137	0.356
SCTNet+L_{cpp}+L_{per}+L_{id}	2.07	0.140	0.362
SCTAdaNet+L_{cpp}	2.29	0.174	0.411
SCTAdaNet+L_{cpp}+L_{id}	2.20	0.163	0.428
SCTAdaNet+L_{cpp}+L_{per}	2.19	0.150	0.405
SCTAdaNet+L_{cpp}+L_{per}+L_{id}	2.29	0.153	0.414

Moving on to the LPIPS metric, it provides valuable insights into the perceptual similarity between the generated frames and the target style. Notably, the SCTNet+L_{cpp}+L_{per} model outperforms the other models in terms of the first frame

(i=1) with a remarkably low LPIPS value of 0.137. This indicates that the proposed model successfully captures and replicates the target style's perceptual features, resulting in a visually convincing first frame.

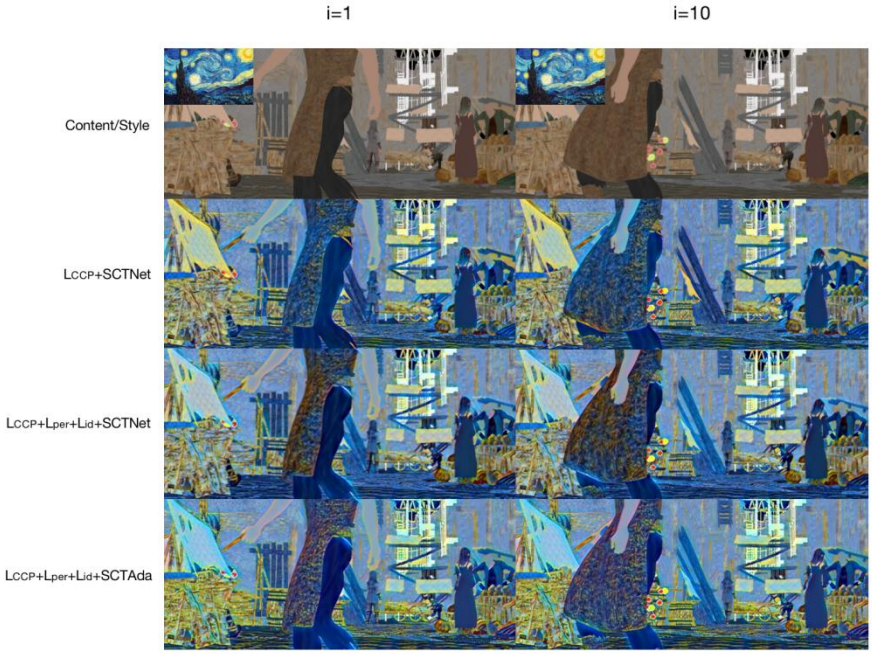

Fig. 4. Visual comparison with video style transfer methods

Overall, the experimental analysis highlights the importance of different loss functions in video style transfer. While the SCTAdaNet model does not demonstrate significant improvements over the SCTNet model in terms of SIFID and LPIPS metrics, the identity loss shows promise in preserving the input video's identity and maintaining style consistency throughout the video sequence. Future research should focus on refining and optimizing the loss functions to further enhance the model's performance.

3.5 Visualization results

In order to qualitatively estimate the migration accuracy, this work also visualized the generation effect of the SCTAdaNet proposed in this article, as shown in Figure 5. The proposed approach enables stable video style transfer with high-quality stylization effects. Image style transfer results across diverse domains and video style

transfer results using various frames from the Sintel dataset are given in the first row and second row, respectively.

Fig. 5. Visualization of transfer results of proposed SCTAdaNet

4 Conclusion

This paper presents a network model for video style transfer that achieves high-quality results by incorporating multiple loss functions and attention mechanisms. Our network architecture, called SCTAda, is an improved version of the SCT network proposed in CCPL. Additionally, this work introduce residual modules and AdaAttN to further enhance the style transfer performance. The loss functions employed in our model include Content Loss L_{cont}, Style Loss L_{sty}, Contrastive Coherence Preserving Loss L_{cpp}, Identity Loss L_{id} and Perceptual Loss L_{per}. Each of them plays a crucial role in guiding the optimization process and improving the quality of generated videos. L_{cpp} enhances coherence and contrast between consecutive frames, promoting visual consistency and artistic coherence. L_{per} combines content and style losses to emphasize perceptual similarity between generated and content videos. L_{id} preserves the identity of the content video, preventing excessive style modification.

While our network model demonstrates promising results in video style transfer, there are potential areas for improvement. First, this work can explore further optimization of the network architecture, such as investigating more complex residual

modules or attention mechanisms to enhance the details and perceptual quality of generated videos. Second, additional constraints and prior knowledge can be considered to exert more control over the style and content of generated videos. Furthermore, exploring diverse style transfer scenarios and real-time video style transfer could be valuable extensions. These advancements would make our network model more practical and adaptable to various applications.

In conclusion, our work contributes to the research and application of video style transfer. However, there are still many directions to explore and improve upon. Through continuous research and innovation, I believe that this work can further advance video style transfer technology, providing users with richer and more personalized visual experiences.

References

1. Gatys, L. A., Ecker, A. S., & Bethge, M. Image Style Transfer Using Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).

2. Johnson, J., Alahi, A., & Li, F. F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. European Conference on Computer Vision (2016).

3. Huang, X., Belongie, S., & Lim, S. N. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. Proceedings of the IEEE International Conference on Computer Vision (2017).

4. Ruder, M., Dosovitskiy, A., & Brox, T. Artistic Style Transfer for Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).

5. Chen, C., Li, X., Li, R., & Tang, J. StyleBank: An Explicit Representation for Neural Image Style Transfer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017).

6. Xing, J., Gao, X., Fan, Q., Yu, D., & Wang, L. Deep Video Style Transfer. Proceedings of the European Conference on Computer Vision (2018).

7. Li, X. et al. CCPL: Contrastive Coherence Preserving Loss for Versatile Style Transfer. In Proceedings of the European Conference on Computer Vision (2018).

8. He, K. et al. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).

9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Advances in Neural Information Processing Systems, pp. 6629-6640 (2017).

10. Zhang, R., Isola, P., & Efros, A. A. Colorful image colorization. In European Conference on Computer Vision, pp. 649-666 (2018).

11. Wang, X., Yu, F., Dou, Q., & Heng, P. A. Video object segmentation with unsupervised learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3231-3239 (2018).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.