Tencent AI Lab

Real-Time Neural Style Transfer for Videos

Haozhi Huang[1,2], Hao Wang[1], Wenhan Luo[1], Lin Ma[1], Wenhao Jiang[1], Xiaolong Zhu[1], Zhifeng Li[1], Wei Liu[1]

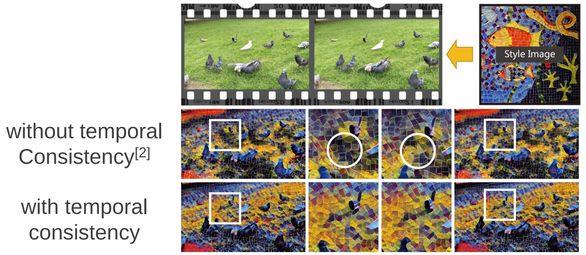
Tencent AI Lab[1] Tsinghua University[2]

Correspondence:
huanghz08@gmail.com
wliu@ee.columbia.edu

CVPR July 21-26 2017 HONOLULU

Introduction

In this paper, we explore the possibility of exploiting a feed-forward neural network to perform style transfer for videos and simultaneously maintain temporal consistency among stylized video frames.



Our key contributions are:

- A novel real-time style transfer method for videos is proposed, which is solely based on a feed-forward convolutional neural network and avoids computing optical flows on the fly.
- We demonstrate that a feed-forward convolutional neural network supervised by a hybrid loss can not only stylize each video frame well, but also maintain the temporal consistency, which is empowered by a proposed two-frame synergic training method.

Previous Methods

Iterative image style transfer[1]

- The first method using deep neural network for image style transfer.
- Based on a very slow iterative optimization.

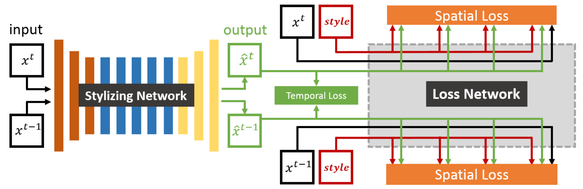
Feed-forward image style transfer[2]

- No more iterative optimization process. Fast.
- No temporal consistency.

Iterative video style transfer[3]

- Incorporate temporal consistency.
- Based on a very slow iterative optimization.

Two-Frame Synergic Training

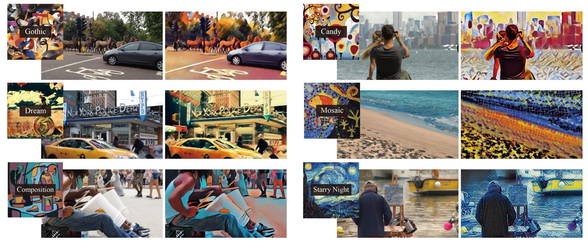


Our style transfer model consists of two parts: a stylizing network and a loss network. The stylizing network takes one frame as input and produces its corresponding stylized output. The loss network, pre-trained on the ImageNet classification task, defines a hybrid loss function including spatial and temporal components.

During the training process, two consecutive input frames x^{t-1} and x^{t} are fed to stylizing network creating two consecutive output frames $\hat{x}^{t-1}$ and $\hat{x}^{t}$. Then the two consecutive output frames are fed to the loss network to compute the hybrid loss. Specifically, the spatial loss is computed separately for each of the two consecutive output frames and the temporal loss is computed based on both of them.

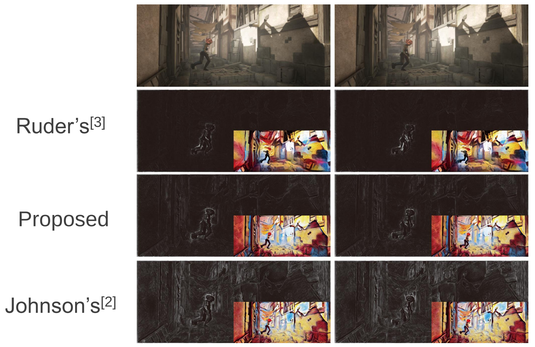
After training, as the stylizing network has already incorporated temporal consistency, we can apply the network frame by frame to an arbitrary input video to generate temporally coherent stylized video.

Some results for different styles are shown below.



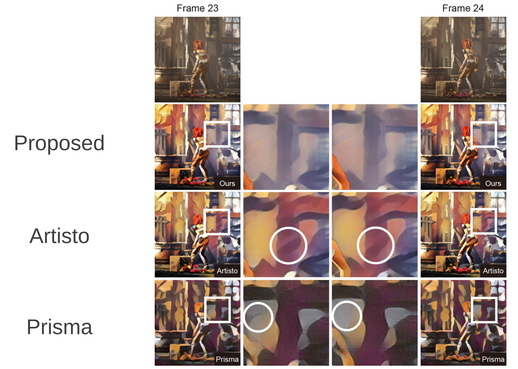
Comparisons

Comparison to previous methods in the literature.



The error maps show the pixel-wise color difference between corresponding pixels of consecutive frames, where the whiter the color is, the larger the error is.

Comparison to commercial softwares.



Check the zoomed areas between consecutive stylized frames. The results suggest that the temporal consistency of our results are better.

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In Proc. CVPR, 2016.
[2] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proc. ECCV, 2016.
[3] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In Proceedings of German Conference on Pattern Recognition, 2016.