HeterStyle: A Heterogeneous Video Style Transfer Application

Xingyu Liu[1], Jingfan Guo[1], Tongwei Ren[1], Yahong Han[2], Lei Huang[1], Gangshan Wu[1]
[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] School of Computer Science and Technology, Tianjin University, Tianjin, China

ABSTRACT

Video style transfer aims to synthesize a stylized video that preserves the content of a given video and is rendered in the style of a reference image. A key issue in video style transfer is how to balance video content preservation and reference style rendering, in order to avoid over-stylization with serious video content loss or under-stylization with unrecognized reference style. In this demonstration, we illustrate a novel video style transfer application, named *HeterStyle*, which can stylize different regions in the video with adaptive intensities. The core algorithm of HeterStyle application is our proposed heterogeneous video style transfer method, which minimizes a heterogeneous style transfer loss function considering content, style and temporal consistency in a Convolutional Neural Networks based optimization framework. With the HeterStyle application, a user can easily generate the stylized videos with good video content preservation and reference style rendering.

CCS CONCEPTS

• **Artificial intelligence** → **Reconstruction**; *Appearance and texture representations*;

KEYWORDS

Video style transfer; heterogeneous stylization; salient object detection; optical flow; convolutional neural networks

1 INTRODUCTION

Video style transfer has gradually attracted public attention in recent years, which aims to synthesize a new video that preserves the content of a given video and presents it in the style of a reference image.[1, 3, 4]. With the help of video style transfer, non-expert users can easily embellish videos and share them with others. However, Most existing video style transfer methods may suffer from over-stylization or under-stylization problems because they use fixed parameters settings during the style transfer(the middle two rows in Figure 1).

In this demonstration, we illustrate a novel video style transfer application, named *HeterStyle*, in order to meet the requirements of balancing video content preservation and reference style rendering. We find that viewers are not attracted to all the content in a content video. They only focus on the attractive content most of time. Hence, we detect the salient regions in the content video and preserve the content of these regions while the other regions can be intensely rendered to emphasize the reference style.

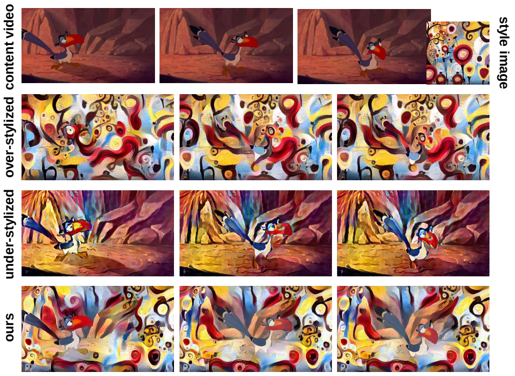

Figure 1: An example of homogeneous video style transfer and heterogeneous video style transfer.

2 HETEROGENEOUS VIDEO STYLE TRANSFER

The core algorithm of HeterStyle application is Heterogeneous Video Style Transfer method. Assume $\boldsymbol{f}$ denotes a frame in a content video, $\boldsymbol{a}$ denotes a style image, and $\boldsymbol{x}$ denotes the corresponding frame of $\boldsymbol{f}$ in the stylized video. Figure 2 shows an overview of our method. Given a content video and a style image, we first estimate both forward and backward optical flow between each pair of the adjacent video frames. Then, we detect the salient object by fusing appearance saliency and motion saliency. Based on the constraint of appearance saliency and optical flow saliency maps, we propose a heterogeneous style transfer loss function consisting of content loss, style loss and temporal consistency loss, and optimize it to generate the final stylized video:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{content} + \beta\mathcal{L}_{style} + \gamma\mathcal{L}_{temporal}, \tag{1}$$

where α, β and γ are weight parameters to control the influences of different losses to the effect of video style transfer.

Content loss. Inspired by [2], we select relu4_2 layer from the VGG-19 network as the content layer. The content loss is calculated based on the weighted squared error between

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10.
https://doi.org/10.1145/3240508.3241402

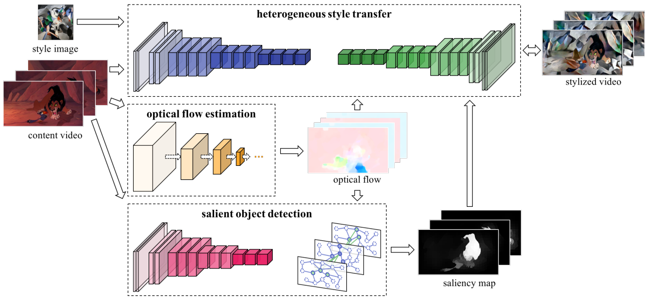

Figure 2: An overview of the proposed method.

$\mathbf{X}^l$ and $\mathbf{F}^l$:

$$\mathcal{L}_{content} = \frac{1}{2}\sum_{i=1}^{M^l}\sum_{j=1}^{N^l}\mathbf{\Omega}_{ij}(\mathbf{X}^l_{ij}-\mathbf{F}^l_{ij})^2, \tag{2}$$

where $\mathbf{\Omega}$ is a weight matrix that is generated from saliency map in the value range of $[0.05, 0.95]$, and resizing it to the size of $M^l \times N^l$.

Style loss. We select relu1_1, relu2_1, relue3_1,relu4_1 layers from the VGG-19 network for style loss calculation. We calculate the Gram matrix of $\mathbf{A}^l$ and $\mathbf{X}^l$, and define the style loss as the weighted sum of the squared errors between these Gram matrices on all the layers:

$$\mathcal{L}_{style} = \sum_{l=1}^{L}\frac{\lambda^l}{(2M^lN^l)^2}\sum_{i=1}^{M^l}\sum_{j=1}^{N^l}(1-\mathbf{\Omega}^l_{ij})\big(\mathcal{G}(\mathbf{X}^l)-\mathcal{G}(\mathbf{A}^l)\big)^2, \tag{3}$$

where $\mathbf{\Omega}^l$ is a weight matrix defined as in Eq. (2); $\mathcal{G}(\cdot)$ denotes the Gram matrix, $\mathcal{G}(\mathbf{X}^l) = \sum_{k=1}^{N^l}\mathbf{X}^l_{ik}\mathbf{X}^l_{jk}$; λ^l is a weight threshold, whose default value is 1; L is the number of layers, which equals 4.

Temporal consistency loss. Similar to [4], we strengthen the temporal consistency in stylized videos with the constraint of optical flow. We detect the disoccluded regions and motion boundaries, and compare the stylized result of the rest regions with the warped previous stylized frame with optical flow:

$$\mathcal{L}_{temporal} = \frac{1}{|\mathcal{H}|}\sum_{p_{ij}\in\mathcal{H}}(\boldsymbol{x}_{ij}-\widetilde{\boldsymbol{x}}_{ij})^2, \tag{4}$$

where $\mathcal{H}$ denotes a set of pixels that do not belong to the disoccluded regions and motion boundaries in $\boldsymbol{f}$; p_{ij} denotes a pixel belonging to $\mathcal{H}$; $\boldsymbol{x}_{ij}$ denotes the stylized result of p_{ij}; $\widetilde{\boldsymbol{x}}$ denotes the warped previous stylized frame with optical flow; $\widetilde{\boldsymbol{x}}_{ij}$ denotes the pixel in the same position to p_{ij} in $\widetilde{\boldsymbol{x}}$; $|\cdot|$ denotes the cardinality of a set.

3 EXPERIMENTS

We compared our method with two state-of-the-art methods: Gatys's method [2] and Ruder's method [4]. To keep the fairness in comparison, we use the same parameter setting for all the methods, *i.e.*, the weights of content loss, style loss and temporal consistency are equal to 1, 20, and 200, respectively.

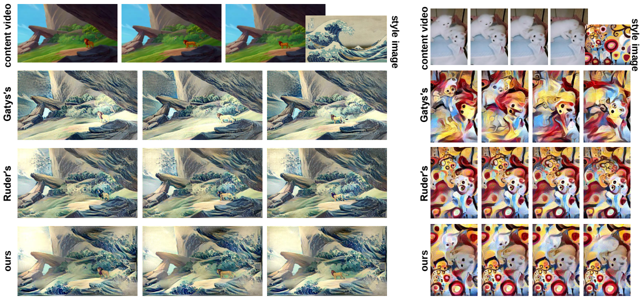

Figure 3: Qualitative examples of stylized videos using different methods.

Figure 3 shows several examples of the stylized videos generated with different methods. We have: 1) Gatys's method generates the stylized videos with obvious jitters because it has no temporal consistency constraint in its optimization function; 2) Ruder's method provides temporally consistent stylized videos, but it may suffer from the problems over-stylization or under-stylization on all or a part of video frames; 3) Our method can handle the videos with various content while using different style images, which balances video content preservation and reference style rendering well.

4 APPLICATION

We develop a mobile phone application, named *HeterStyle*, based on the proposed heterogeneous video style transfer method. Figure 4 shows the procedure of using HeterStyle application. With the HeterStyle application, a user only needs to shoot a video, select a style, and then enjoy the generated stylized video.

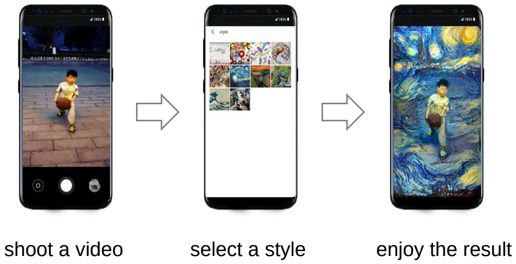

Figure 4: The procedure of using HeterStyle.

5 CONCLUSION

In this demonstration, we illustrated the HeterStyle application, which can preserve the video content well in the salient regions while emphasizing the reference style by rendering the inconspicuous regions intensely. The experimental results showed that HeterStyle application can help a user to easily generate the stylized videos with good video content preservation and reference style rendering.

REFERENCES

[1] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. 2017. Coherent Online Video Style Transfer. In *IEEE International Conference on Computer Vision*. IEEE, 1105–1114.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2414–2423.

[3] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. 2017. Real-time neural style transfer for videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7044–7052.

[4] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. 2016. Artistic style transfer for videos. In *German Conference on Pattern Recognition*. Springer, 26–36.