Photorealistic Style Transfer for Videos

Michael Honke, Rahul Iyer, Dishant Mittal
University of Waterloo

Introduction

Photo-realistic style transfer is a technique which transfers colour from one reference domain to another domain by using deep learning and optimization techniques. Here, we present a technique which we use to transfer style and colour from a reference image to a video. We take help from 2 different papers.

Deep Photo Style Transfer Luan et al.

- They portrayed a deep-learning technique particularly for style transfer in photorealistic images that was able to control a huge diversity of image content while reliably transplanting the style from a reference image.
- The styles are transferred by segmenting the image and computing losses between different segments (eg. sky segment in reference to sky segment in content).
- They utilised the Matting Laplacian to restrict the conversion from the input to the output to be locally affine in colourspace.

Artistic style transfer for videos Ruder et al.

- They portrayed a technique which enables the transfer of style from an artistic image (for instance, a Van Gogh painting) to a complete video sequence.
- They deduced that independently processing each frame of the video leads to glimmering and untrue irregularities.
- Using the optical flow, pixels that are consistent between frames have a high temporal loss applied to them, forcing them to be consistent. Recently disoccluded pixels have reduced or zero temporal loss weighting.

Loss Functions

Style Loss

$$Loss_m = \sum_{c=1}^{3} V_c[O]^T M_I V_c[O]$$

where M_I is a standard linear system (Matting Laplacian) that only depends on the input image I and $V_c[O]$ is the vectorized version (N 1) of the output image O.

$$Loss^l_{s+} = \sum_{c=1}^{C} \frac{1}{2N^2_{l,c}} \sum_{i,j} (G_{l,c}[O] - G_{l,c}[S])^2_{i,j}$$

where G is the Gram matrix and S is the style image.

Video Loss

$$\mathcal{L}_{video}(f^{(i)}, a, x^{(i)}) = \alpha \mathcal{L}_{content}(f^{(i)}, x^{(i)}) + \beta \mathcal{L}_{style}(a, x^{(i)}) + \gamma \sum_{j \epsilon J, (i-j) \geq 1} \mathcal{L}_{temporal}(x^{(i)}, w^i_{i-j}(x^{i-j}), c^{(i-j,i)})$$

- i denotes the index of a frame
- $f^{(i)}$ is the i^{th} frame of the video
- a is the style image
- $x^{(i)}$ is the i^{th} stylized frame to be generated
- c denotes temporal weight
- w is a function that warps a given frame using the optical flow field that was estimated between two images
- J denotes the set of indices each frame should take into account relative to the frame number, e.g., with J = 1, 2, 4
- $\mathcal{L}_{content}$ and $\mathcal{L}_{style}$ are defined by Gatys et al. [3]

Merged Loss Function

To implement the style transfer from one photorealistic image to a whole video sequence using deep learning we merged the concepts from both the works i.e from Ruder et al. [1] and Luan et al. [2]. Precisely, we accomplished this in two parts:

- We integrate the the loss functions used in style transfer in photorealistic images [2] to the loss functions used in artistic style transfer in videos [1].
- We used the semantic segmentation technique proposed by Zhou et al. This helped us to automate the task of segmentation which can then be used in computing the overall loss for the videos. The style loss term in artistic style transfer in videos [1] is replaced completely with the semantic segmentation loss used by Luan et al. [2]. The final Loss function can be written as:

$$\mathcal{L}_{final} = \sum_{1=1}^{L} \alpha_l \mathcal{L}^l_c + \tau \sum_{1=1}^{L} \beta_l \mathcal{L}^l_{s+} + \lambda \mathcal{L}_m + \gamma \sum_{j \epsilon J, (i-j) \geq 1} \mathcal{L}_{temporal}(x^{(i)}, w^i_{i-j}(x^{i-j}), c^{(i-j,i)}_{long})$$

, where the symbols are as defined in the previous sections.

Method

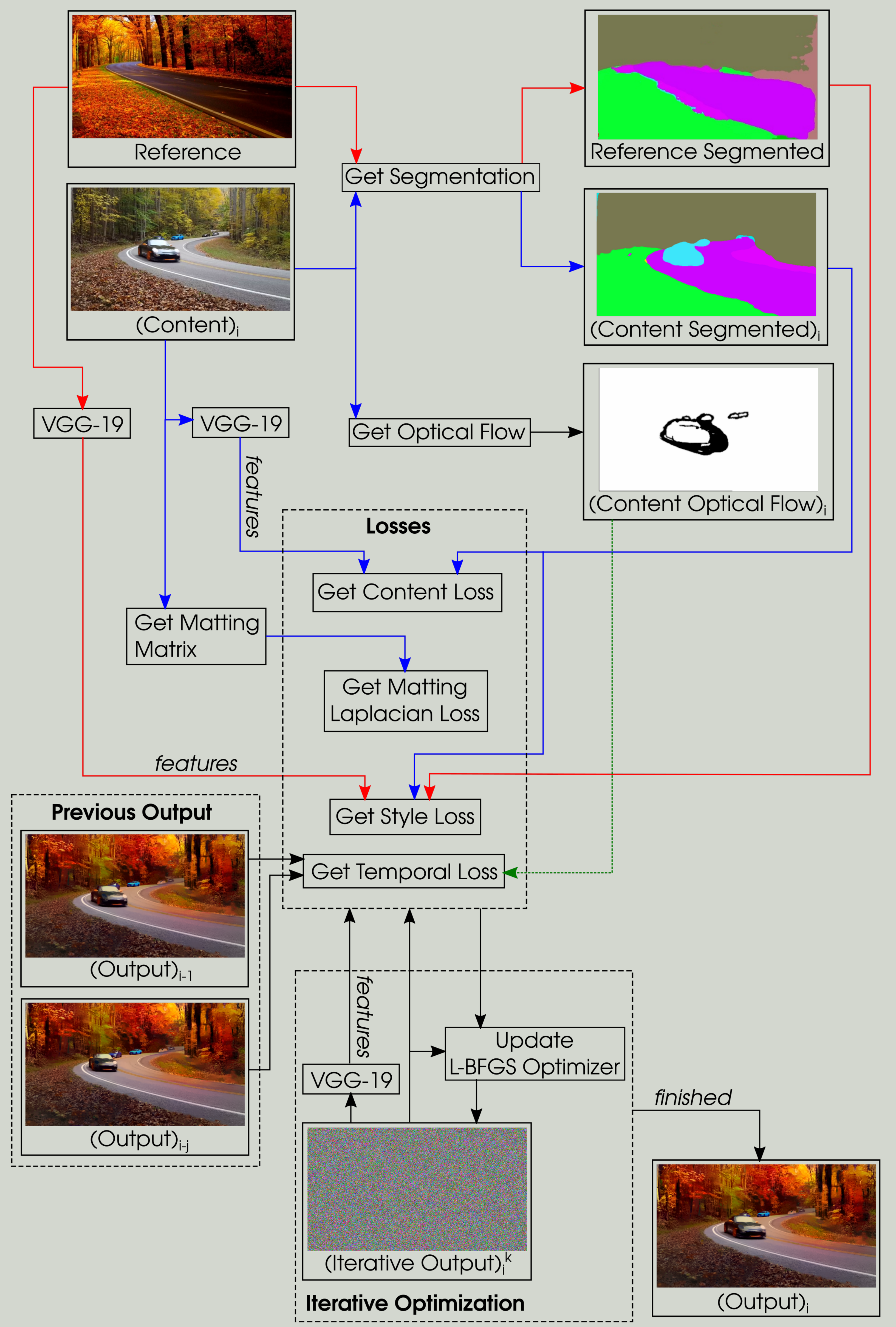
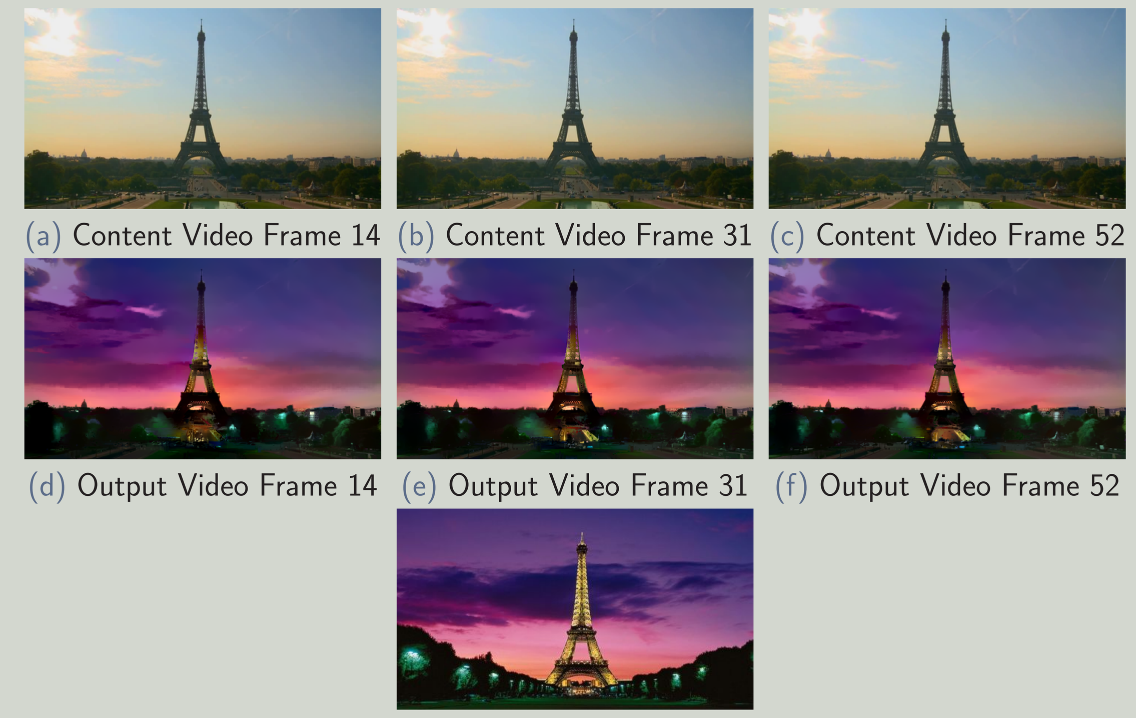


Figure: Technique

Results

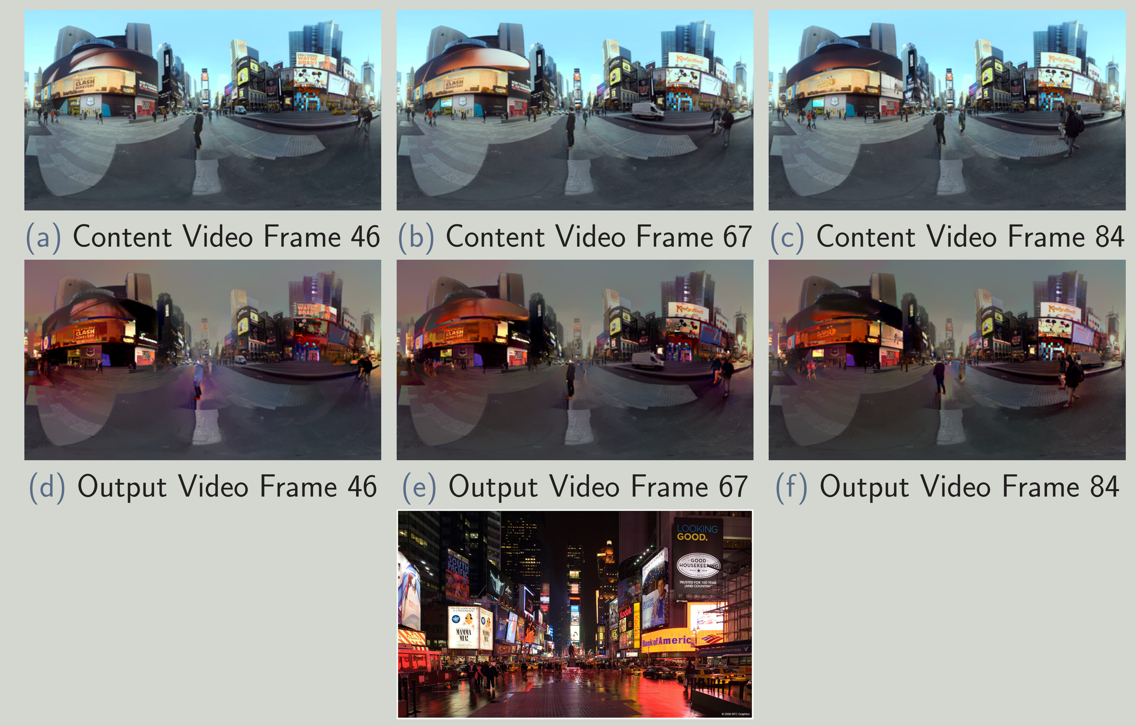
This section shows the results generated by applying the photorealistic video style transfer technique to 3 different videos along with their corresponding reference frames.

Figure: Eiffel Tower Photorealistic Video Style Transfer Results



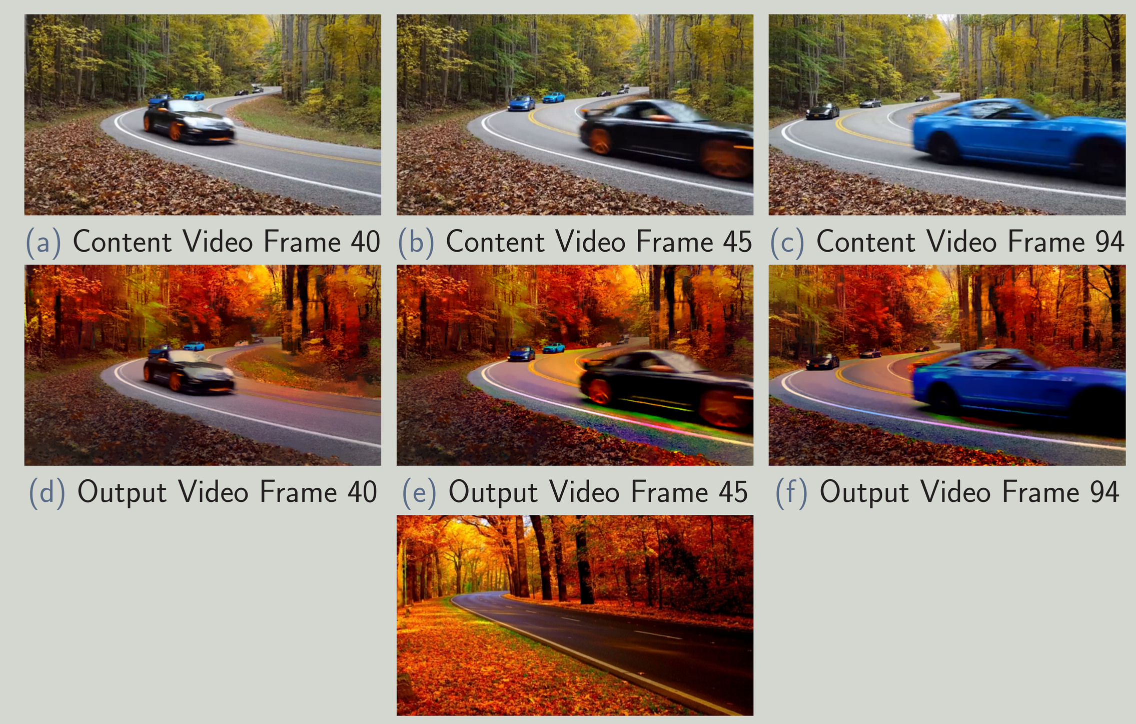
(a) Content Video Frame 14 (b) Content Video Frame 31 (c) Content Video Frame 52 (d) Output Video Frame 14 (e) Output Video Frame 31 (f) Output Video Frame 52 (g) Reference image

Figure: New York Time Square Photorealistic Video Style Transfer Results



(a) Content Video Frame 46 (b) Content Video Frame 67 (c) Content Video Frame 84 (d) Output Video Frame 46 (e) Output Video Frame 67 (f) Output Video Frame 84 (g) Reference image

Figure: Moving Cars Photorealistic Video Style Transfer Results



(a) Content Video Frame 40 (b) Content Video Frame 45 (c) Content Video Frame 94 (d) Output Video Frame 40 (e) Output Video Frame 45 (f) Output Video Frame 94 (g) Reference image

Conclusion

- We produced a novel combination, bringing photorealistic style transfer to video content.
- We demonstrate that Ruder et. al.'s [1] video style transfer algorithm is both robust and adaptable enough to be used in new style transfer applications.
- We further validate Luan et al.'s research [2] by showing that it is consistent enough to be applied to multiple video frames.

[1] Ruder, M., Dosovitskiy, A., & Brox, T. (2016, September). Artistic style transfer for videos. In German Conference on Pattern Recognition (pp. 26-36). Springer International Publishing.
[2] Luan, F., Paris, S., Shechtman, E., & Bala, K. (2017). Deep Photo Style Transfer. arXiv preprint arXiv:1703.07511.
[3] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2414-2423).