Stylistic Music Video Through Neural Matting and Style Transfer*

*ECSE 544- Final Project, April 13, 2023

Mohamed Debbagh
Dept. Bioresource Engineering
McGill University
Montreal, Canada
mohamed.debbagh@mail.mcgill.ca

***Index Terms*—style transfer, video matting, video processing, deep learning**

I. Introduction

Video styling and special effects are important components to developing a popular music video. Music artists often accompany their songs with visuals to showcase their own styles outside of the auditory experience. Similar to taste and smell, audio and visual stimuli are often linked together and tend to enhance general feelings conveyed by the music, and vice-versa. While the story-line of a music video is not as significant to the experience, the visual effects will often make or break a song. Songs with music videos tend to exponentially increase an artist's visibility within the industry. While established artists spend millions of dollars on sets and special effects, smaller underground artists rely on abstract and innovative software visualization techniques to enhance their music videos and stand out from the crowd. With significant advances in deep learning, many of these effects can be achieved through computational photography and video processing techniques. This paper explores two techniques that can enhance a music video, these include re-styling of the video using abstract visual style transfer and video matting to achieve a separation of foreground and background similar to a green screen. Applying these visual techniques on videos are often more complicated than standard single image processing since they need to maintain temporal coherence between frames and must be considered throughout. To showcase these effects, these techniques are applied to a generic rap music video [1]. Both techniques applied in this paper utilize neural network models and thus we consider them neural methods.

II. Video Matting

The first technique implemented in this project is image compositing and matting. Image matting is a fundamental challenge explored in computational photography and describes the process of separating the foreground and background within an image. While traditional movie production methods implement image compositing through green screens and early matting techniques such as rotoscopy, novel deep learning based approaches have shown promising results. These approaches learn to capture complex features and patterns within an image, enabling them to handle a wide range of scenarios and produce high-quality alpha mattes. Models such as Deep Image Matting, MODNet and U^2-Net [2]–[4] have demonstrated state-of-the-art results outperforming traditional methods in both controlled and natural scenes(non-green screen). This allows small productions, such as the one described in the scenario of this project, to take videos from any scene location within a small budget and transform them into creative environments. In this project we implement U^2-Net as our deep network to produce alpha-mattes.

A. U^2-Net

U^2-Net is a Fully Convolutional Network (FCN) architecture developed by Qin et al. for the task of salient object detection (SOD), and particularly for the task of image matting. It is composed of a two layered nested U-net like architecture that captures and refines the hierarchical features of an image. The innovative component the architecture is the use of Residual U-blocks (RSU), which enable the network to capture multi-level scales of features effectively. This allows U^2-Net to produce accurate and detailed image mattes, even in challenging scenarios with complex foregrounds and backgrounds. A pre-trained U^2-Net model trained on the DUTS-TR data-set of 10533 images for salient object detection, was provided by the original authors of the U^2-Net paper and was used for the task of producing the alpha matte masks. samples of the masks are shown in Fig. 1.

B. Implementation

Since the goal of this paper is to perform compositing over the whole video, individual frames of the scene are fed as input to the U^2-Net model to produce a corresponding alpha-matte. Once the original frames and resulting matte's were produced and properly labeled, a simple alpha blending operation is applied using the python implementation of the openCV library [5]. Alpha blending is the operation of overlaying a

Fig. 1. mask samples produced by U^2-Net. Original frames(left), Alpha matte mask(right).

foreground onto a new background using the alpha matte. the process equation is as follows,

$$I = +(1-\alpha)B$$

where, I is the resultant image, F is the foreground image, B is the background image, and α is the alpha matte. This operation is performed over all frames of the original video and applied to a new background selected by the author of this paper. See Fig. 2 for an example of the resulting operation on a single frame.

Fig. 2. Alpha blending operation performed on one of the frames of the original video. Original frames(bottom left), Alpha matte (top left), New background frame (top right), Resultant frame from alpha blending operation (bottom left).

III. Style Transfer

The stylizing of a video is one of the more expressive parts of visual effects. Often times abstract expressions borrowed from specific art styles are desired for a particular scene such as in a music video. One method explored in the field of computational photography that is used to translate abstract art styles onto images is known as style transfer. Style transfer techniques combine the content of one image with the artistic style of another, known as the style reference. The resulting image is a visually appealing output that retains the structure and content of the original image while incorporating the distinct artistic characteristics of the style image. Style transfer has become a popular topic, as significant advancements in deep learning techniques have achieved impressive results.

A. Neural Style Transfer

Neural Style Transfer (NST) is a technique introduced by Gatys et al. in 2015 that employs a pre-trained Convolutional Neural Network (CNN) architecture, typically VGG-19, to extract feature representations from both the content and style reference [6]. The objective of the NST is to minimize the content loss, which measures the difference between the content image and the generated image, and the style loss, which measures the difference in style between the style reference and the generated image. The combination of these two losses results in a visually appealing output that successfully merges the content and style of the input images. While this approach has led to compelling results, the style transfer process is inefficient as the training step is performed each time when applying styles to new images. This results in a computationally inefficient process. This is especially problematic when performing style transfer on videos which require performing NST on many frames. Many variants of NST have been proposed to address this limitation. One such method is known as arbitrary style transfer, which employs adaptive instance normalization to apply style transfer without having to retrain the model [7]. Golnaz et al. further expands upon this concept by proposing Conditional Instance Normalization (CIN) method, which enables real-time style transfer using a single feed-forward network [8]. This method is desirable for video processing as it is a computationally efficient NST method. Thus this method was selected to perform NST for this paper. A TensorFlow implementation [9] of this method exists and was employed to perform rapid NST on each frame of the music video.

B. Temporal Consistency

Temporal consistency is an important consideration for video style transfer, as it ensures a smooth and coherent stylization across all frames in the video sequence. When NST is applied to each frame independently, the stylization results in inconsistencies across frames such as flickering and other artifacts. Various techniques, such as incorporating optical flow, can be used to ensure smooth transitions and consistent style application throughout the video sequence. However a simpler method is employed in this paper as a robust solution was not necessary for the case of this simple music video. Temporal consistency was achieved by blending the previous stylized frame, referred to as the ghost frame, with the current content frame. The blending process starts by calculating a weighted average of the current content frame and the ghost frame. The weights are determined by a transparency factor, which can be adjusted to control the relative contribution of each frame to the blended result. A higher transparency factor gives more weight to the ghost frame, while a lower transparency factor prioritizes the content frame. This weighted blending ensures that the stylized output of the current frame is influenced by the previous frame's stylization, allowing

for a smoother transition between adjacent frames. While the simplicity of this method works for the applications of this paper, there are better and more robust approaches to temporal coherence in NST in videos including, optical flow-based warping, temporal loss functions, and temporal smoothing. The code implementation of this method is provided by [10]. See Fig. 3 for output from the NST network used in this paper.

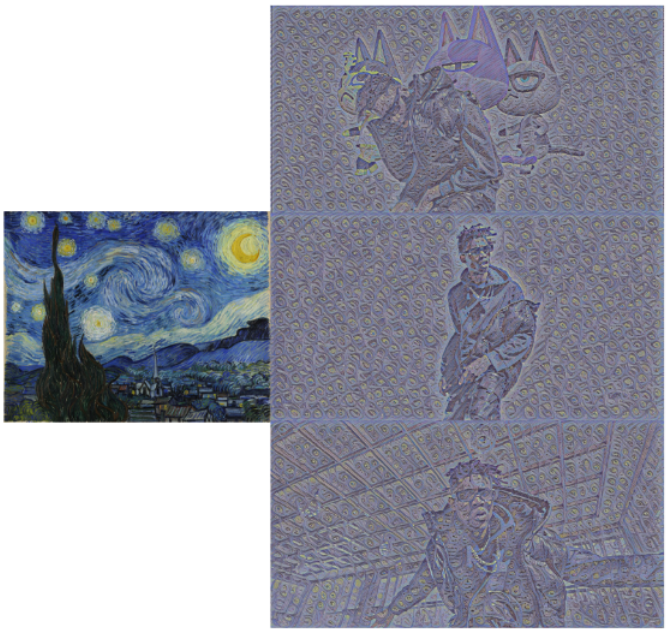

Fig. 3. Style transfer applied to various frames throughout the music video. Reference style(left), stylized output frames(right)

IV. Discussion

The U^2-Net model was able to generate alpha mattes for the rap music video with a fairly good accuracy. The alpha mattes were applied to the video frames using alpha blending to create a new background. The new background that was applied to the original foreground was creative and unique and enhanced the overall mood and tone of the video. In general, since the image matting was applied on natural images, not all alpha mattes produced perfect separation. Further post-processing should be done to correct these errors.

The proposed arbitrary style transfer method was able to achieve real-time stylization for each frame of the video. The stylized video reproduced the visual style of Van gogh's *Starry Night* that was consistent throughout the entire video. While the final stylized video was visually impressive, some flickering and discoloration was observed in certain frames due to using a simplistic temporal coherence method. Future work could include exploring more robust methods that incorporate optical flow and a smoothening loss function to ensure smoother transitions between frames.

Overall, the combination of video matting and style transfer greatly enhanced the visual appeal of the rap music video and demonstrates the potential of neural network-based video processing techniques in creative applications.

V. Conclusion

Recent developments in deep learning methods have made is possible to produce powerful visual effects for even for small low budget productions to produce impressive videos. This paper explored two computational techniques used to enhance a music video through image compositing and matting, and style transfer. The implementation of these techniques utilized neural network models, making them highly effective. Image matting was achieved using U^2-Net, which was able to produce accurate and detailed alpha mattes, allowing for compositing over any scene location within a small budget. This paper also explored a real-time style transfer technique using a single feed-forward network and an arbitrary stylization method desirable for video processing. Additionally, the paper highlighted the importance of temporal consistency for video style transfer, which was achieved through blending the previous stylized frame with the current content frame. While the simplicity of the method used in this paper works for its application, more robust approaches such as optical flow-based warping, temporal loss functions, and temporal smoothing are available for better temporal coherence in style transfer for videos. Overall, this paper demonstrated the potential of computational photography and video processing techniques to enhance the visual effects of music videos and enable smaller productions to achieve creative environments at a lower budget.

References

[1] "A rapper man free stock video footage, royalty-free 4k amp; hd video clip," May 2021. [Online]. Available: https://www.pexels.com/video/a-rapper-man-7955111/

[2] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," 2017.

[3] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," in *AAAI*, 2022.

[4] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," vol. 106, 2020, p. 107404.

[5] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[6] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015.

[7] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[8] G. Ghiasi, H. Lee, M. Kudlur, V. Dumoulin, and J. Shlens, "Exploring the structure of a real-time, arbitrary neural artistic stylization network," *arXiv preprint arXiv:1705.06830*, 2017.

[9] T. Hub, "Magenta: Arbitrary image stylization," https://tfhub.dev/google/magenta/arbitrary-image-stylization-v1, accessed: 2023-04-05.

[10] B. Westgarth and T. Jogminas, "style-transfer-video-processor," https://github.com/westgarthb/style-transfer-video-processor, 2021.