On Extending Image Style Transfer to Video Style Transfer

Di Wang
Carnegie Mellon University
Pittsburgh, PA 15213
diwang2@andrew.cmu.edu

Hiroshi Wu
Carnegie Mellon University
Pittsburgh, PA 15213
bw1@andrew.cmu.edu

Alvin Pan
Carnegie Mellon University
Pittsburgh, PA 15213
qpan@andrew.cmu.edu

1 Introduction and Problem Setup

Neural networks are a fundamental building block of many machine learning applications. Expanding on neural networks' abilities, [1] presents a convincing method of generating a new image from an old image that preserves content, but captures the "style" of another style image. To date, most neural style transfer literature presents innovations to image-to-image style transfer, but the topic of video-to-video style transfer from a single style image remains relatively untouched. Video style transfer can be naively done through image style transfer done on every frame of the video. However, this yields poor results from frame-to-frame coherency, as defined below. Since two neighboring frames in the video are processed entirely separately, it is possible for the resulting frames to have differing textures even in locations that are supposed to be continuous from one frame to another. This results in a great amount of flickering, which is very unpleasing to the eye and greatly downgrades the quality of the video.

We investigate methods to fix this flickering issue through added coherence methods, and attempt to show acceptable results when compared to the naive method of video-to-video style transfer.

We consider two facets of this problem. First, we explore some ways to quantify the notion of "coherence" and analyze its results. Second, we utilize some techniques to enforce frame coherence in our current models. This is an area we intend to explore further, which will be detailed in Section 7. For now, we add an additional term to our loss function based off of a simple definition of frame-to-frame coherence.

Throughout this paper, our video representation is a sequence of frames, or images of the same size. For our preliminary result, we define frame-to-frame coherence measure as the mean-squared error between each pixel value of two consecutive frames. we partition each video into scenes, where a scene is a sequence of contiguous frames such that the mean-squared error is relatively small over each pairwise consecutive frames in the entire scene. We also define scene alignment to be a condition in which a consecutive sequence of frames is a subset of a sequence of frames that consist of a scene.

2 Literature Review

The natural way to attempt the problem is to use a frame-by-frame approach based on image-to-image neural style transfer. However, from [3] and our experimental results, a naive application of [1] produces flickering and false discontinuities, which [3] attempts to solve by introducing a temporal consistency constraint based on optical flow, which encourages frame-to-frame consistency as well as discourage the formation of dis-occluded regions, or near blank regions. However, the algorithm has only been applied to videos that span only one scene.

[2] introduces a sequence of reference frames, which are a sequence of photos that correspond to frames in the video, but in the desired style. This method produces high-quality results, both in terms of realism image quality, and the features transferred, but it is inherently niche, as the existence of a

Preprint. Under review.

sequence of coinciding and similar frames of a different style does not apply to the general problem of video-to-video neural style transfer from a style image.

3 Data

Style images are selected based off of how far their style differs from our content videos. As such, our style images are selected from art, especially images of paintings by painters like Vincent van Gogh and Pablo Picasso to observe what modern videos look like after being transferred into various art styles.

Our content video dataset currently consists of mostly animations, so our results may not be representative of the whole corpus of possible video style transfers. However, in the current stage we only investigate the basics of the viability of coherent video style transfer, so therefore currently we will use this limited dataset. We hope to expand upon this in the future.

4 Methods and models

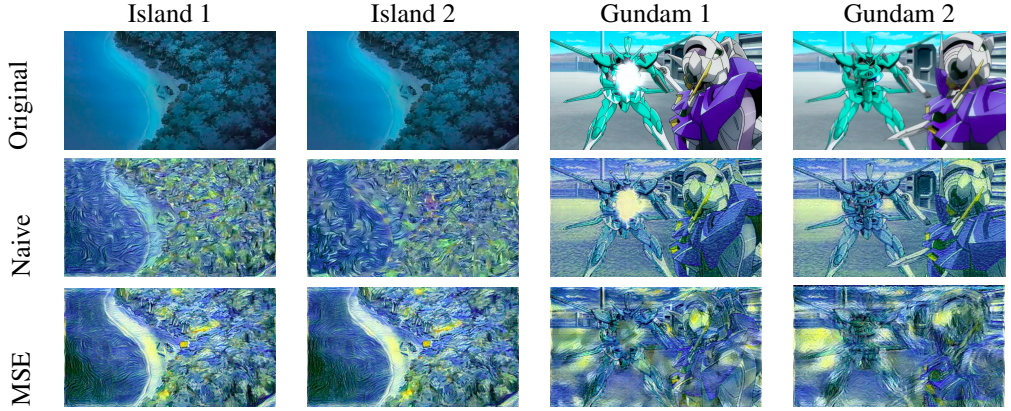
Our baseline model uses [1] for our naive frame-by-frame style transfer. For our improvement, we add our coherence metric to the objective function from [1], and allow the neural network to optimize over the metric as well to ensure frame-by-frame coherency.

Our modified loss function for each frame with coherency is as follows:

$$L_{total}(\vec{p_i}, \vec{a}, \vec{x_i}) = \alpha L_{content}(\vec{p_i}, \vec{x_i}) + \beta L_{style}(\vec{a_i}, \vec{x_i}) + \eta L_{coherency}(\vec{x_i}, \vec{x_{i-1}})$$

where $\vec{p_i}$ is the i-th content image vector, $\vec{a_i}$ is the i-th style image vector, $\vec{x_i}$ is the i-th image that is the image generated by the model and α and β are weighting factors for content and style reconstructions respectively. The content loss and style loss functions are those as defined in [1]. Our coherency loss term is the mean-squared-error between the pixel values of the previous frame and the current frame. As such, we are operating under the assumption that the video in its entirety will be scene aligned, so all frames in the video are part of one scene.

5 Preliminary Results



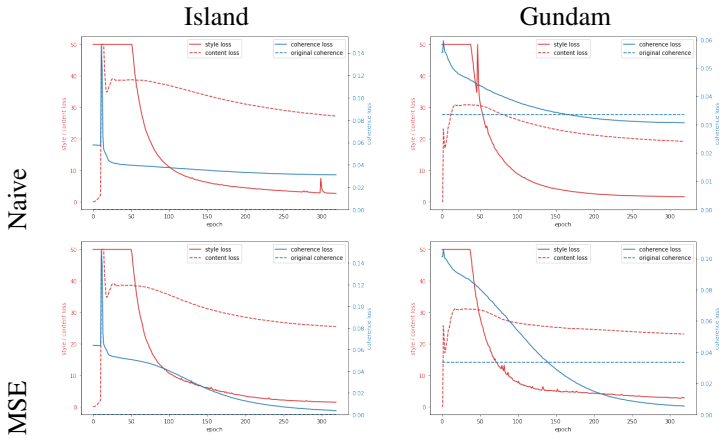
For both videos, [video name] 1 and [video name] 2 denote two consecutive frames from the original video, the naive method video, and the naive method with coherency video, respectively. Style image was a facsimile of Vincent Van Gogh's Starry Night.

For the naive model video, low coherency is clearly observable in the common case. An example is the Island frame transition in the table. The transition without the coherence metric shows drastic changes on the coastline over the two frames, while the two frames trained with the coherence metric optimized over the change of the previous frame. However, when we apply our new model to a video that violates scene alignment, severe ghosting is observed as the penalty for the coherence dominates all other optimization terms in our loss function. The gundam transition perfectly justifies this, where

ghosting is observed during the frame transition due to the white flash of the bullet shot by the green robot. In this case, the naive model actually works a lot better under this circumstance since we hardly observe any ghosting or shadowing, as it is not operating under any scene alignment assumption.

However, the common case for any two frames is to be in a scene, so in general, the naive method with the coherency constraint performs much better than the model without the coherency constraint.

6 Evaluation of preliminary work



When the scene alignment is satisfied, the naive method with the added coherency constraint performs much better. This is shown by the island example and its corresponding average training graph for one frame. Note that the original coherency loss within a scene is very low. Hence, an addition of the coherency loss to the gradient smooths the transition between two frames and helps encourage the neural network to generate a new frame that stays within the same scene in the generated video.

When it is not satisfied, the naive method will outperform the one with added coherency in the case when scene alignment is not satisfied. This is shown by the gundam example and the loss graph for the gundam example as well. Here, the addition of a coherency loss to the gradient omits certain content features and causes ghosting (due to the inclusion of some features from the previous frame) of the original work in an effort to optimize over coherency.

7 Future work

Our models would be optimized towards the common case on most videos. It is very rare for a scene transition to occur within a video. However, any scene transition will cause artifacts to all the frames after that particular scene change, which we hope to mitigate with the inclusion of a second-order delta-based loss term. In addition, our definition of coherence is defined mostly qualitative without any rigorous mathematical proof. Hence, we intend to rigorously show why our metric works to improve coherency or improve upon our current metric.

We intend to augment our dataset by including videos from a wide range of films and animations, to ensure varying characteristics such as drastic scene changes, rate of bright flashes, etc., and test the generality of our model.

8 Teammates and work division

Di will write most of the reports and provide theoretical basis for our results, in addition to debugging. Hiroshi will handle most of the programming, namely coding the models and setting up the training pipeline. Alvin will prepare the datasets, handle the media necessities of the project, and fine-tune the model hyperparameters.

References

[1] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[2] Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. Stylizing video by example. *ACM Transactions on Graphics*, 38(4), 2019.

[3] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. *CoRR*, abs/1604.08610, 2016.