Photo-realistic Style Transfer for Videos

Manoj Ravi
manojr@stanford.edu

Naveen Mohan
nav3en@stanford.edu

1. Abstract

This paper focuses on applying deep-learning techniques to perform style transfer from a photographic image to a video in a photo-realistic way. Transfer learning using neural networks for images has been gaining momentum and more attention since Gatys et al.[1] published the results of merging the style component from artistic images with the content of real-world photographs. Their approach utilized higher-level feature representation of images from hidden layers of VGG convolutional network. This work was followed up by artistic style transfer for videos by Manuel et al.[2] where a temporal constraint was used to regularize the style and content loss components and create better stability between frames when transferring style for videos.However, these two methods were more conducive for painterly or artistic style transformations and less for photo-realistic style transfer. So Fujun et al.[3] introduced a constraint that controlled the transformation from the input to the output to be locally affine in color space and expressed this constraint as a custom fully differentiable energy term. This methodology has helped achieve certain level of photo-realistic style transfer for images using a concept of Matting Laplacian. Our approach builds on this by trying to apply the approaches from the deep photo style transfer paper to small videos to overcome drawbacks of applying the artistic style transfer techniques to stylize videos. We show that the style transfer can be achieved with good results for few small videos and also compare the results observed with results obtaining using Cycle GANs to transfer style.

2. Introduction

The world of transfer learning uses the knowledge of pre-trained models to utilize it as an initializing point of action or feature extractor. The decisions on how to use existing pre-trained models depend on nature of the new dataset for which we have to accommodate the training and testing phases. Depending on the nature of the new dataset and the size, we could decide if we need to fine-tune the parameters or not. There are primary aspects that decide the restrictions for using pre-trained CNNs like that of manipulating the architecture and using minimal learning rates. Transfer learning gives us scope to deal with the scenarios by leveraging the already existing labeled data of some related task or domain. We try to store this knowledge gained in solving the actual task and extend it to our problem of interest. Transfer learning involves the handling of domain and task. Domain defines the features to be handled and the task defines the labels that need to be handled based on marginal probability distribution.

There are multiple domains where transfer learning gets used, some of the major ones include: learning from simulation, learning from other domains, learning underlying structure of images, learning domain invariant representations, etc.

1.1 Problem Statement

In this paper we explore how to extend the neural style transfer concept to videos with the original video not losing transfer faithfulness at the same time maintaining the content of the original video without distortion and spill over, particularly for use cases where we do not have enough data for transfer learning. Transferring content and style from two different sources and balancing it right proportion is a work in progress, particularly for photo-realistic transfers and making it work for videos without losing the information in consequent frames is a challenge. In the course of experimenting these aspects for videos we would also like to understand how the different components of a frame and a video interact with neural networks and understand how different features get extracted and learnt and multiple layers within the model.

1.2 Data set

We are planning to use sample images and videos from the following two datasets:

a) YouTube 8 million dataset for the videos (https://research.google.com/youtube8m/)

b) Google Open Image dataset for the images (https://github.com/openimages/dataset)

Since we are leveraging transfer learning, we will need fewer input images and videos. For one complete run we would just need one content video and a style image.
For Cycle GANs since we are using the model weights that are pre-trained we don't need additional data for training the GANs.

1.3 Evaluation Methodology

We will be evaluating this project in two broad ways:

Qualitative (User based feedback on style transfer efficiency and elimination of distortion): Here, we will check how good the stylized videos look and how do their presentation in terms of distortion and spill over look. We will evaluate these in light of the faithful transfer of the realistic style to the video.

Quantitative (We are planning to use the results from artistic style transfer for images and videos as the baseline. The efficiency for the new implementation will be evaluated with respect to the baseline. We hope to arrive at a loss function or a nuanced implementation, which will help us achieve better performance compared to the baseline.

Since there is no way of directly comparing the loses across different videos and there is no global way to measure style transfer in a quantified way, we will mostly be relying on qualitative methods to evaluate the results.

Figure 1: Baseline Image Results a) Input Image b) Reference Image c) Output Image

3. Related Work

Artistic Style Transfer for images
There has been a lot of work done with artistic style transfer for images and it has been gathering a lot of academic and industry attention because of the numerous applications [1] [5] [6]. There has been constant work on improving the style transfer algorithm to make it more efficient and converge faster [7] [8] . Several papers focus on exploring and improving different aspects of the style transfer process like improving the stability and specific characteristics of the images produced [9] [10] [11]

Artistic Style Transfer for videos and Optical Flow
The algorithm for style transfer for images has been extended for videos by incorporating the concept of optical flow loss [2].
Lot of work has focused on trying to use optical flow information to augment the style transfer to prevent occlusion and blurring effects when porting style transfer from images to videos [12] [13]

Photographic Style Transfer for images
Most recently the style transfer algorithm was augmented to address various challenges when applying style transfer for realistic photographs with great success [3]. This incorporated work on Matting Laplacians [14] and semantic segmentation. Semantic segmentation which has also seen a lot of interest [15] has also been something that has been used to aide the style transfer process [16]

Generative Adversarial Networks[17]:

Generative Adversarial Nets (GANs) is a framework paradigm that trains two models: a) a generative model that captures the data distribution b) a discriminative model that captures the probability if a sample data came from training data or G. GANs have been very successful in image generation and representation learning. GANs are evaluated based on a loss function that forces images from G to be as closely representative of the original image thus trying to maximize the probability of error by model D.

Image to Image Translation Paradigms:

This entire project could be paraphrased as an image-to-image translation effort as well if we consider the videos as individual frames. We have these translation efforts right from image analogies, which utilizes non-parametric ways of conversion. Later efforts tried to build on deep learning models like that of CNNs to generate parametric functions. There are even Bayesian frameworks based on Markovian priors that help in adopting styles from multiple images. There is also a scope of work which constrains the input and output to share content features and forces the output to be in a input related space like image pixel or image feature space [17]. There are other similar methodologies like collection style transfer, object transfiguration and photo enhancement, which handle unpaired images as well but are not discussed in this paper.

CycleGANs extend these ideas from traditional image translation methods and from deep neural networks to generate the models D and G without any paired images/ extensive data sets to train the model. Thus it makes the objective function not specific to dataset types and more general purpose in nature. The results from Cycle GANs have been found to be much promising in comparison to some of the existing parallel frameworks like BiGAN, CoGAN,etc.

4. Methodology:

We utilize a pre-built VGG-Network [18] as the basis for performing style transfer. We use the weights from the first 16 layers to generate the features to be used for the style transfer. The main aim is to generate a stylized image x based on the contents of original image p and the style from another image a. We define content and style loses and formulate the following objective function as defined by Gatys et. all in the paper [1]

$$\mathcal{L}_{singleimage}(\boldsymbol{p},\boldsymbol{a},\boldsymbol{x}) = \alpha\mathcal{L}_{content}(\boldsymbol{p},\boldsymbol{x}) + \beta\mathcal{L}_{style}(\boldsymbol{a},\boldsymbol{x})$$

Equation 1: artistic style transfer loss term

Where the style and content loss are given by:

$$\mathcal{L}_{content}(\boldsymbol{p},\boldsymbol{x}) = \sum_{l\in L_{content}} \frac{1}{N_l M_l} \sum_{i,j} \left(F_{ij}^l - P_{ij}^l\right)^2$$

Equation 2: content loss term

The content loss measures the mean squared distance between the pixels of the original image and the generated stylized output image

$$\mathcal{L}_{style}(\boldsymbol{a},\boldsymbol{x}) = \sum_{l\in L_{style}} \frac{1}{N_l^2 M_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2$$

Equation 3: style loss term

The style loss measures the mean squared distance between the Gram Matrices of the style or reference image and the stylized output image. The Gram matrices give a good representation of which features activate together in an image and this helps us capture the style and textures of images in a better manner.
An additional optical flow loss term was added to the total loss to extend the style transfer to video.

$$\mathcal{L}_{temporal}(\boldsymbol{x},\boldsymbol{\omega},\boldsymbol{c}) = \frac{1}{D}\sum_{k=1}^{D} c_k \cdot (x_k - \omega_k)^2$$

Equation 4: temporal loss term

The temporal loss term penalizes regions in the output where optical flow is consistent but the warped image deviates from it. We can calculate the temporal loss over short and long term. Long term penalizes deviations from more than just one previous frame.

For overcoming challenges with spillover effect and local distortions, the deep photo paper introduced a photorealism regularization term to constrain the transformation to be locally affine in the color space.

$$\mathcal{L}_m = \sum_{c=1}^{3} V_c[O]^T \mathcal{M}_I V_c[O]$$

Equation 5: photorealism regularization term

This is achieved using calculating a matting Laplacian matrix M and using that to calculate the photorealism regularization term.

In addition semantic segmentation is also used to generate masks for the content and style images to constrain the style transfer and to avoid spill over effects. The semantic segment masks are passed as channels along with the input video frame images and the style loss is calculated using this.

The plan is to study how these techniques can be leveraged effectively in order to apply photo realistic transfer for videos. The following loss equation reflects the intention.

The total loss is the sum of content, style, temporal/optical losses and the photorealistic regularization term.

$$L_{total} = \sum_{l=1}^{L} \alpha_l L_c^l + \Gamma \sum_{l=1}^{L} \beta_l L_s^l + \tau \sum_{l=1}^{L} \gamma_l L(sh,lo)_t^l + \lambda L_m$$

Equation 6: total photo-realistic vide style loss term

Using this as the objective function we use Gradient descent to minimize the loss and update the pixels of the output video frame based on the direction of the negative

gradients. This helps us effectively transfer the style from the reference image to the input frames of the video and finally produce the resultant output frame.

Once we achieve this we explore the use of Cycle GANs (Generative Adversarial Networks) [17] to see if we can extend the photo-realistic style transfer logic for images to videos and achieve results on par with or better than what we achieved earlier.

The CycleGAN architecture is an adaptation of the neural style transfer architecture [19]. If X and Y are the two domains between which the model needs to learn the mapping. There is a forward function and a backward function to be learnt on top of couple of discriminator functions.

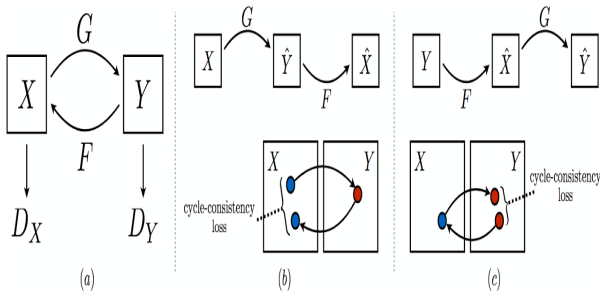


Figure2 : CycleGAN modeling: mapping X to Y and inverse using G, F functions while optimizing on discriminator functions [19]

There are two types of losses that need to be handled in CycleGANs - a) adversarial loss b) cycle consistency loss. Objective Function:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F),$$

$$G^*, F^* = \arg\min_{F,G} \max_{D_x, D_Y} \mathcal{L}(G, F, D_X, D_Y).$$

Equation 7: a) Final Objective Function b) Optimization Function [19]

As we could see the final objective function on which the CycleGAN is built on loss functions which try to work generating images similar to input images, while trying to identify the translated samples against original samples (adversarial loss).

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))],$$

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].$$

Equation 8: Adversarial and Cyclic consistency loss [19]

But as we could see, the GAN loss (adversarial loss) results in the possibility of a given input image to get mapped to a randomized permutation of images [19]. So the paper introduces cycle consistency loss factor which reduces the possible mapping space.

We need to achieve,

$$x \rightarrow G(x) \rightarrow F(G(x)) \approx x;$$
$$y \rightarrow F(y) \rightarrow G(F(y)) \approx y$$

this ensures consistency of forward and backward loss (per figure) where F is the reverse mapping function from Y->X. [19].

Though the nature of losses calculated here are not comparable with that of neural style losses in the physical world, the outputs produced could be qualitatively compared to see how reference images customize the content and styles. We will try to gauge that for a pair of images and also utilize some pre-trained weights to guide that translation.

As part of our experiments, we try to optimize the results for (neural style transfer for videos + deep photo) using hyper parameter tuning and then try to compare the results from the customized algorithm (neural style transfer for videos + deep photo) to CycleGANs qualitatively to see how it fares.

5 Experimentation and Results

We started by experimenting with the artistic style transfer for pictures. Figure 3 shows the results of transferring artistic style to a photo.

Figure 3: Artistic Style Transfer for Image Results a) Input Image b) Style Image c) Output Image

Here we observe that the style from the style image has been applied quite effectively to the input picture.

Next we looked at artistic style transfer for videos, which transferred a style from a picture to a video. This extends the style transfer for images by incorporating a temporal loss to maintain consistency over the frames in the video. Figure 3 outlines the results we observed.

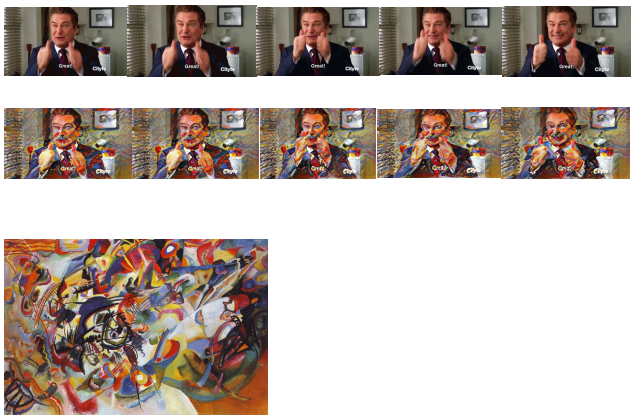


Figure 4: Artistic Style Transfer for Videos Results a) Frames of Input Video b) Frames of Output Stylized Video c) Style Image

The figure shows a few frames from the original video and the output video. We observe that the style from the style image has been applied to the content video.

Following this we worked on applying the artistic video style transfer algorithm on a video using a realistic photo as the style image. This would serve as a good baseline to evaluate the results of applying the final algorithm. Figure 4 contains the results we observed.

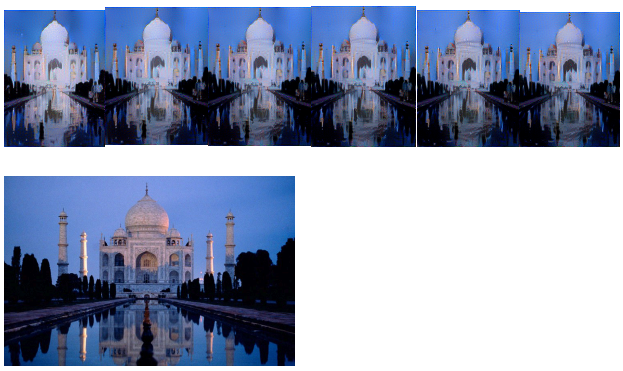


Figure 5: Artistic Style Transfer for Videos Results a) Frames of Input Video b) Frames of Output Stylized Video c) Style Photo

As seen in the images above, the style is not completely transferred. We observe inconsistencies and distortions in the frames of the video.
The results obtained from applying the artistic style transfer for videos for transferring style from realistic photos can be used as a baseline. We look to improve upon by addressing shortcomings of this approach for realistic photos.

Similar results were observed when we applied the style transfer for realistic images. The deep photo style transfer aims to address these challenges by using semantic segmentation and using a photorealistic regularization term. The results of using the deep photo style transfer algorithm on the realistic images resulted in promising results which Figure 5 displays

Figure 6: Artistic Style Transfer for Videos Results a) Input Image b) Style Photo c) Output during training d) Final output

As observed, this algorithm produces really good and realistic results for photo realistic style transfer. There are lesser number of distortions and inconsistencies.

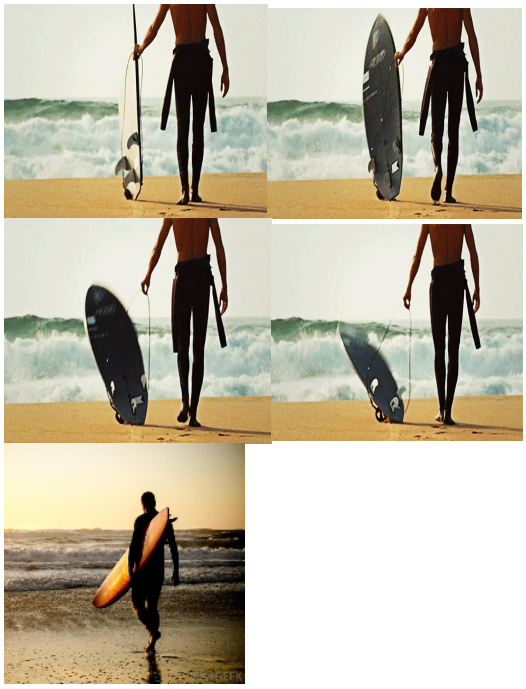


Figure 7: a) Input video frames b) Reference style image

We implemented the final algorithm in two stages. The first stage involved including the photorealism regularization term and then followed by including the semantic segmentation part. We outline the experiments across the two stages. We worked mainly with 200 x 200 pixel size for the videos and the reference photos.
The algorithm has a lot of hyperparameters which can be tuned, starting with the weights – α, β, γ and λ. We experimented with different values of the weights. The values of α and β had a impact on the tradeoff between the content and style image features in the ouput image[FIGURE]. The value of γ controlled the temporal loss and ensured the smoothness of the output video. The best values for these changed from one video to another. But we observed the photorealism regularization parameter value $\lambda = 10000$ gave the best value across different videos we experimented with.

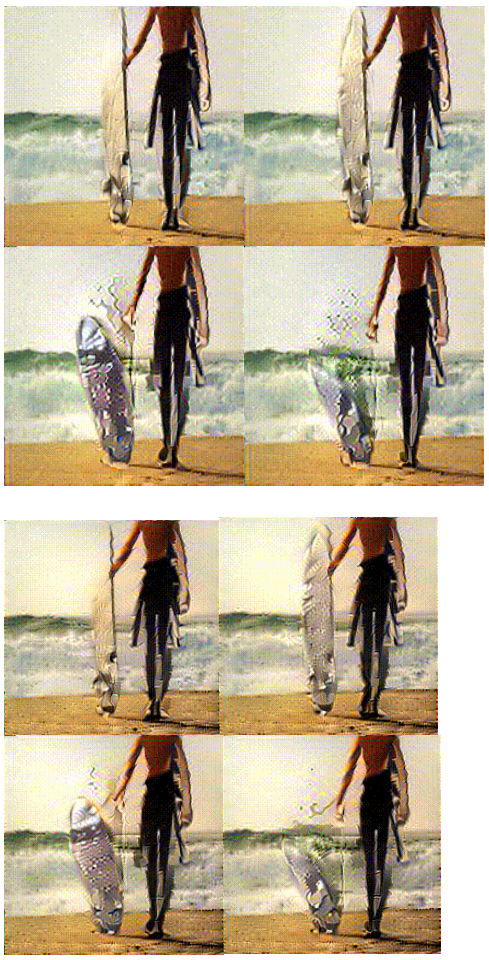
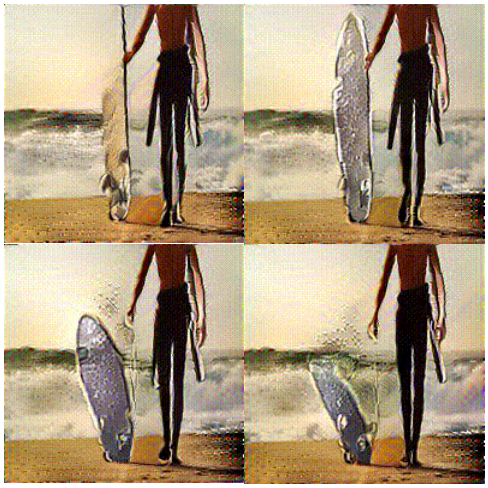


Figure 8: Results with different images experimenting with different alpha and values
a) Surfing without Segmentation - Different values of hyperparameters – alpha = 20, Beta = 20, gamma = 100, lamda = 10000
b) Surfing without Segmentation
- With different parameter values alpha = 5, Beta = 100, gamma = 200, lamda = 10000

In addition to these we also experimented with using weights from different layers of VGG-Network to calculate the content and style loss. We experiment with a few combinations but that didn't have too much of a positive impact on the output. There is potential to experiment with more combinations. For majority of the results observed we used conv4_2 for the content loss and conv1_1, conv2_1, conv3_1 and conv4_1 for the style loss

Figure 9 – Results obtained without segmentation

After including segmentation we ran more experiments mainly with the different weight values and outlines the results observed. Including segmentation had a positive impact in videos where the segments were clearly distinguishable as expected. Using semantic segmentation helped address some of the spill over issues that was prevalent in the output video frames without segmentation.



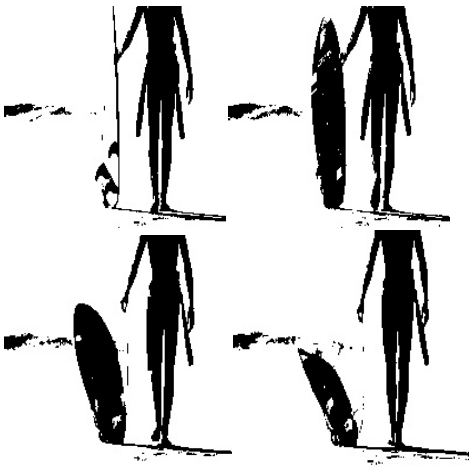


Figure 10 – Results obtained after segmentation
a) Surfing with Segmentation - With different parameter values alpha = 5, Beta = 100, gamma = 200, lambda = 10000
b) Masks for the input video frames

We observed that the outputs produced by algorithm successfully transferred style from the reference to the frames of the input video. However there is still more room for improving the temporal loss component in order to improve the quality of the video.

CycleGAN results:

We approach CycleGANs in the same way we approached the custom neural style transfer and we tend to get results where mapping of components of a given frame are much more aligned and focused with very minimal hyperparameter tuning. In figure below, we see the results of translating a surfer video from summer to winter based on a pre-trained model, which captures nature components during summer and winter and has understood the translation.

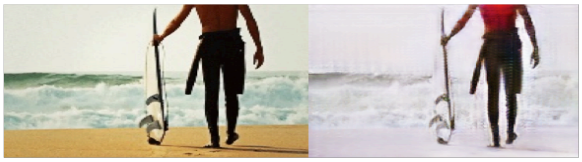


Figure 11: CycleGAN translation of an image of a surfer in a beach from summer to winter with snow

We extend the concept of domain translation (nature images) with the task being constant (conversion from summer to winter) here. The nature of the pre-trained weights was such that they had learnt image translation from summer to winter during day times only. After minimal hyper-parameter tuning of the lambda factor and

without any segmentation we were able to achieve the results in the above figure (video gif attached).

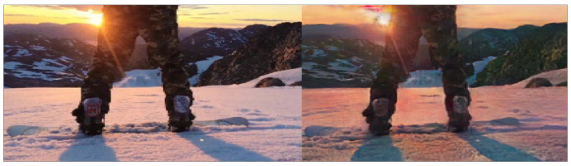


Figure 12: CycleGAN translation of an image of a snow boarder on mountaintop from winter with snow to summer

Then, the same concept was applied to convert a video of a snowboarder during twilight and we attempted to convert that video to that of a summer video. Even after ample effort on parameter tuning, the best result we could obtain was the one below (video gif attached). We could see that the effort to convert the white snow to grass is observed in the output (below the snowboard) but the model's knowledge of the only day time translation restricts its efficacy to videos similar to that of the surfer in the previous examples. This exposes the limitation of the CycleGAN architecture and areas where there is scope for further improvement particularly in the combined area of both domain and task translation.

6 Future Works

As mentioned earlier there is still a lot of room for experimentation with different hyperparameter values. We can experiment with different types of optical flow loss short term and long term flow loss and also trying out other optical flow estimation algorithms other than Deep Flow. There is also potential to study the effects of using different layers in a more detailed manner. One other key area that can be worth looking into is using other semantic segmentation algorithms to come up with better technique to detect and mask the major objects in the frames. We can also try to use more advanced object detection algorithms other than semantic segmentation to guide the style transfer. There are numerous possibilities to explore from the perspective of GANs. One option is to look into integrating loss from Cycle GAN and neural style transfer We can also look into hyperparameter tuning for Cycle GANs and experiment with using other types of GANs and techniques like Pix2Pix etc.

7 Conclusions

We presented a technique to transfer the style from reference photos to reference images in a realistic manner by proposing a combined loss function that ensures style is transferred in a photo-realistic way while maintaining the temporal consistency. We compared the results obtained using the transfer learning approach to a results obtained using Cycle GANs. While Cycle GANs look really promising they still can't help perform style across different domains. They require a lot of data to train models, which are restricted to the domain in which they are trained on. The traditional approach on the other hands are more flexible and help in producing realistically stylized videos from just using a single additional reference image. There is still a lot of potential to explore different avenues to improve the realistic style transfer to videos.

7 Acknowledgements

We'd like to acknowledge the two projects we built upon and based our implementation upon the neural style transfer repo - https://github.com/cysmith/neural-style-tf and the deep photo repo - https://github.com/martinbenson/deep-photo-styletransfer

References

[1] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. CoRR abs/1508.06576 (2015), http://arxiv.org/abs/1508.06576

[2] Manuel Ruder, Alexey Dosovitskiy, Thomas Brox,: Artistic style transfer for videos, https://arxiv.org/abs/1604.08610

[3] Fujun Luan,Sylvain Paris,Eli Shechtman, Kavita Bala: Deep Photo Style Transfer, https://arxiv.org/abs/1703.07511

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros : Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, https://arxiv.org/pdf/1703.10593.pdf

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016.

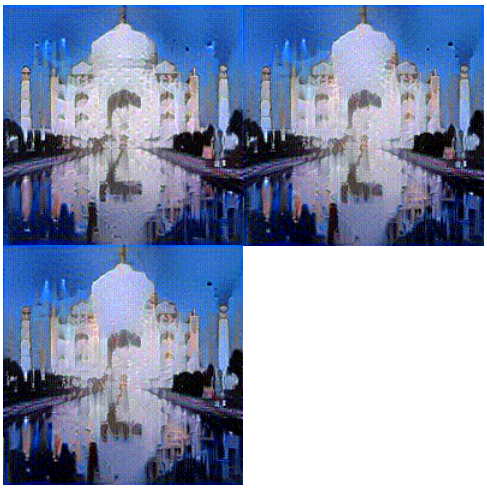
[6] Demystifying Neural Style Transfer Yanghao Li† Naiyan Wang‡ Jiaying Liu† Xiaodi Hou Institute of Computer Science and Techology, Peking University ‡ TuSimple https://arxiv.org/pdf/1701.01036v1.pdf

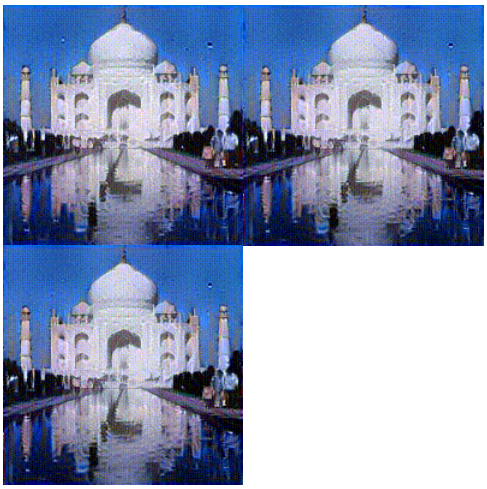
[7] A Learned Representation For Artistic Style Vincent Dumoulin, Jonathon Shlens, Manjunath Kudlur https://arxiv.org/abs/1610.07629

[8] Perceptual Losses for Real-Time Style Transfer and Super-Resolution Justin Johnson, Alexandre Alahi, Li Fei-Fei
https://arxiv.org/abs/1603.08155
[9] Nikulin, Y., Novak, R.: Exploring the neural algorithm of artistic style. CoRR abs/1602.07188 (2016), http://arxiv.org/abs/1602.07188
[10] Characterizing and Improving Stability in Neural Style Transfer
Agrim Gupta, Justin Johnson, Alexandre Alahi, and Li Fei-Fei Department of Computer Science, Stanford University
https://arxiv.org/pdf/1705.02092v1.pdf
[11] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. arXiv preprint arXiv:1611.07865, 2016. 1
[12] FlowNet: Learning Optical Flow with Convolutional Networks Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox,
https://arxiv.org/abs/1504.06852
[13] DeepFlow: Large displacement optical flow with deep matching Philippe Weinzaepfel Jerome Revaud Zaid Harchaoui Cordelia Schmid
https://pdfs.semanticscholar.org/cad6/bf3a625be27cd3e2b2130ba18a04d2aa9cea.pdf
[14] A. Levin, D. Lischinski and Y. Weiss, "A Closed-Form Solution to Natural Image Matting," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 228-242, Feb. 2008, http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4359322&isnumber=4407426
[15] DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille
https://arxiv.org/abs/1606.00915
[16] Semantic style transfer and turn- ing two-bit doodles into fine artworks. A. J. Champandard. arXiv preprint arXiv:1603.01768, 2016
[17] Generative Adversarial Networks Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio
https://arxiv.org/abs/1406.2661
[18] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] (2014). URL http://arxiv.org/abs/1409. 1556. ArXiv: 1409.1556
[19] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros : Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,
https://arxiv.org/pdf/1703.10593.pdf

APPENDIX

Quick overview of a different example of the other photo-realistic video style transfers results. As you can see the values of photorealism regularization term lambda and segmentation have a key effect on the quality of the output video. These help in removing distortions and avoiding spill over effects.





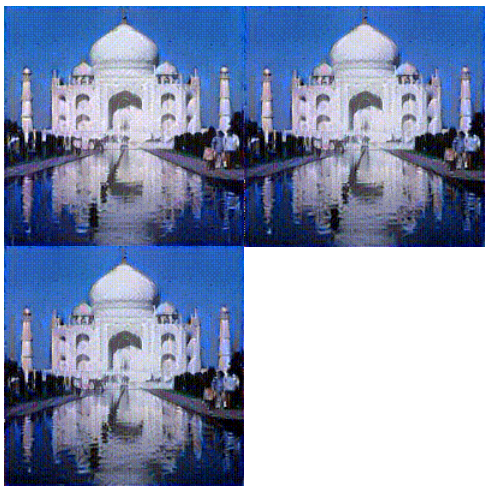


Figure 13: - a) Outputs before using segmentation - Lamda value = 1000
b) Outputs before using segmentation - Lamda value = 10000
c) Outputs after using segmentation - Lamda value = 10000