International Journal on Applications in Engineering and Technology
Volume 8: Issue 1 : June2022 , pp 1 – 7 www.aetsjournal.com ISSN (Online) : 2455-0523

IMPLEMENTATION OF SEMANTIC STYLE TRANSFER BASED ON NEURAL NETWORKS

S.ASWIN JEYA SELVAN , C.ILAYA SUDHIR ,V.GOPI KRISHNA ,
M.MUKESH KRISHNAN

***Abstract*—** Developing an artwork by means of human handwork requires a lot of time and extraordinarythinking capabilities. Each of the artists hastheir own idea of creating artwork.Creating an artwork is not an easy task. Hence we developed a semantic style transfer technique based on Mask R-CNN,one can easily able to create an artwork with the help of content image and the style image.This technique uses segmentation on the content image and content based selection from the animated images to maintain improved semantic features in the resultant image.Our results show that this technique preserves more semantic information in the content image than previous other techniques, while still successfully transferring the art style of a particular style image through texture and color, especially in the background. This technique also effectively avoid semantic mismatch in the processof image style transfer. That is, it can manage semantic consistency in the process of style transfer.

***Keywords*—**: Image style transfer, Semantic segmentation, Semantic Mismatching, Mask R-CNN.

I. INTRODUCTION

In the growing field of artificial intelligence in computer science, the question of whether artificial intelligence and neural networks can produce art has been very much of interest. Styletransfertechnique has been recently taken the artificial intelligence field by storm. The paper "A Neural Algorithm of Artistic Style"[2] hasbeen enforced many times, and these implementations have the ability to transfer style of a certain painting, such as Van Gogh's "Starry Night" or Edvard Munch's "The Scream" to a photographic artwork. In the animation industry, creating the artwork for animated movies is a meticulous and time-consuming process. It is important for the movies to maintain a coherent art style across all frames. This can be especially sensitive when rendering detailed backgrounds, such as countryside landscapes, cityscapes, and Skyscraper buildings.

Style transfer is especially tough for photographs with rigid lines, such as buildings and monuments.

Previous implementations of style transfer methods have numerous limitations, such as transferring the style of animated images to detailed landscapes and transferring of animated structures to buildings. The semantic meaning of a content image is lost because the objects blend together, losing structural, color, and other visual information frequently.

Our ultimate goal is to overcome these limitations by segmenting content images and identifying multiple style images to maintain the semantics of the content image and to capturing the essence of a certain movies or films in a semantically segmented way across images with the newly developed style transfer approach.

II. RELATEDWORK

The original style transfer methodology using neural networks was proposed by Gatys et al [2] in " A Neural Algorithm of Artistic Style ", which performs grade ascent on a white noise image, minimizing two losses content loss against a content image and style loss against a style image. The style loss function is taken at multiple layers in the network, and characterized by the mean squared loss between the Gram matrices of the input image and the artwork. Since also, style transfer has been a extensively explored content, and numerous have worked towards advancements on the algorithm. Patch- grounded styles have been developed using CNN features, similar as in the Neural Doodles

S. Aswin Jeya Selvan , Department of Computer Science and Engineering , Francis Xavier Engineering College ,Tamilnadu , India .(Email : aswinjeyaselvans.ug18.cs@francisxavier.ac.in)

C.Ilaya Sudhir , Department of Computer Science and Engineering , Francis Xavier Engineering College ,Tamilnadu , India .(Email ilayasudhirc.ug18.cs@francisxavier.ac.in)

V.Gopi Krishna , Department of Computer Science and Engineering , Francis Xavier Engineering College ,Tamilnadu , India .(Email : gopikrishnav.ug18.cs@francisxavier.ac.in)

M.Mukesh Krishnan , Assistant Professor , Department of Computer Science and Engineering , Francis Xavier Engineering College ,Tamilnadu , India .(Email : mukesh@francisxavier.ac.in)

project,[7] which used manually authored pixel markers as semantic annotations, achieving advanced quality results than vanilla style transfer. also, a recent paper by Gatys, etal.[3] anatomized different ways in controlling color, spatial position, and scale. They used guided Gram matrices and guided totalities in controlling spatial position. Results showed notable advancements in maintaining semantic meaning in respects to foreground and background in the content image. In the same tone, as the exploration done by Gatys, etal., Li, etal. combined Markov Random Fields with Convolutional Neural Networks[18] to more really transfer semantically analogous corridor of a style image over to a content image. Small workshop has been done on Semantic style transfer specifically for animated styles. still, an open source interpretation of Studio Ghibli's Toonz software, dubbed OpenToonz, was lately released. OpenToonz advertises the capability to perform anime- suchlike style transfer, although they don't reveal the system they use.

III. TECHNICAL APPROACH

1) Dataset

We collected frames from various animated images from google Image Search for our style images. For our content images, we collected pictures ofvarious landscapes and buildings from Google Image search.

2) Baseline

For our baseline, we applied simple style transfer on content images of various portraits and style images of various animated style images.

3) Segmentation Approaches

We firstly planned to apply object spotting and localization to transfer style from object to object across images. still, we set up that nuanced segmentation is delicate indeed with more recent, advanced ways; for illustration, Liang, etal.[8] set up that the CRF- RNN model had the topmost performance in terms of delicacy among FCN, DeepLab, and DeconvNet, so we tested segmentation on Zheng etal.'s perpetration of CRF-RNN[9]. Unfortunately, because of the limited orders, the results were poor for our geography-type images, which didn't have easily defined objects like bikes, boats, or motorcars. therefore, we rather determined that our semantics would be secerning from structures in the focus to geography in the background. Below we show our results after trial with simpler styles to utmost nearly member the focus from the background.

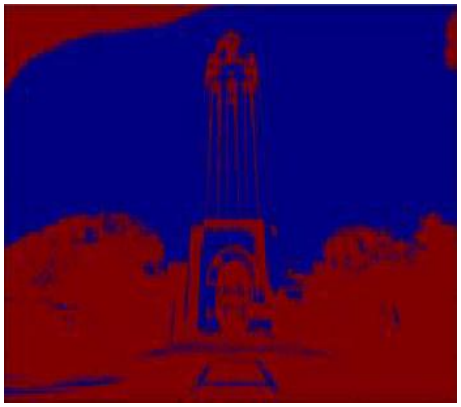
A. Binary segmentation using Otsu thresholding

Otsu's technique, named after Nobuyuki Otsu, automatically performs clustering- grounded image thresholding to reduce a grayscale image to a binary image [8]. The algorithm assumes that there are two classes of pixels following abi-modal histogram and also calculates the optimal threshold similar that theirinter-class friction is greatest. We converted our input image to grayscale and also performed Otsu's methodology through the scikit- image package [11] but found that just in the Gates image(Fig. 1 [b][ii]) the structure had been easily distinguished from the background.

a) Content Images

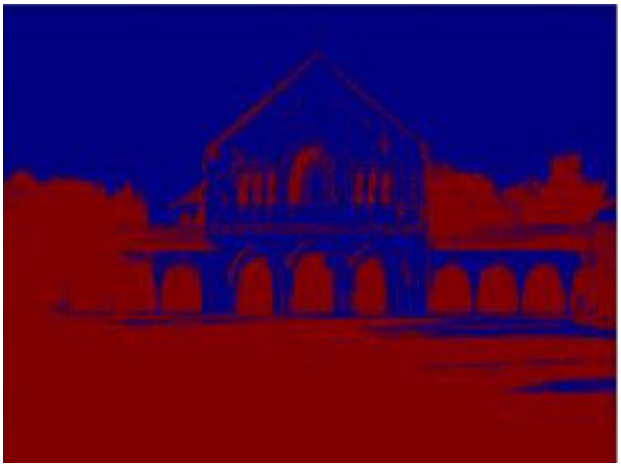
(i)

(ii)



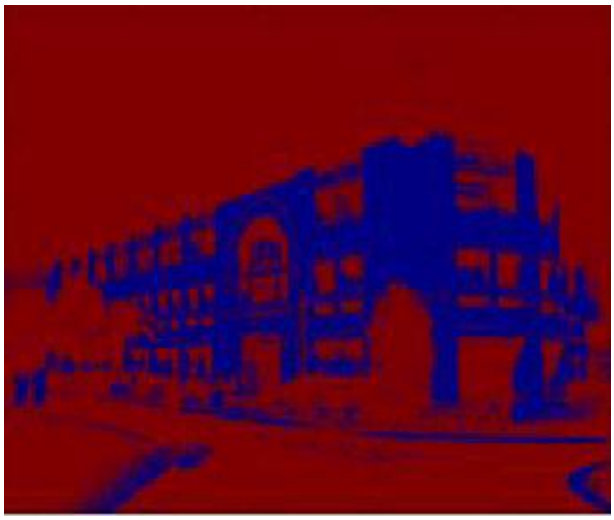
(i)

(iii)



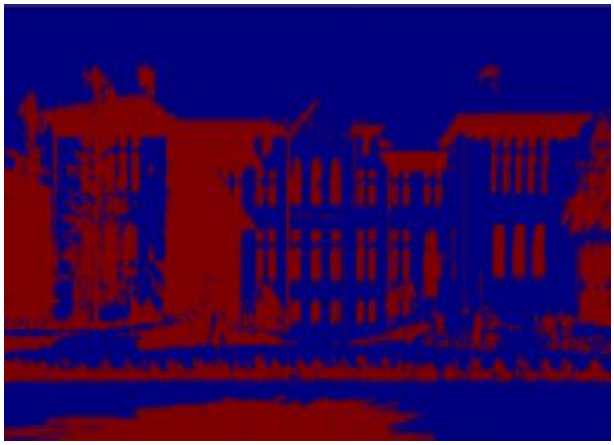
(ii)

(iv)



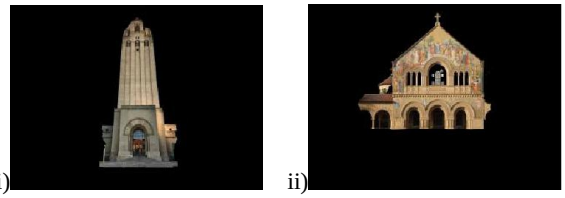
(iii)

B. Segmentation results using Otsu thresholding



(iv)

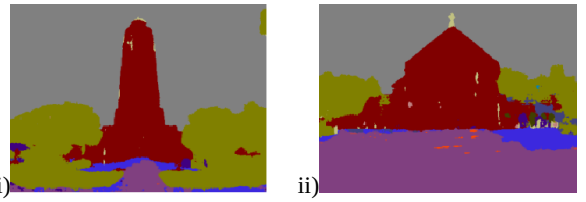
C. Results using the GrabCut algorithm



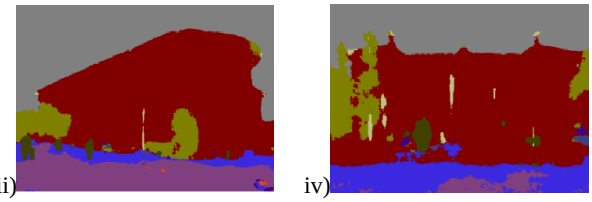
i) ii)

iii) iv)

D. Results using SegNet



i) ii)



iii) iv)

Figure1. Segmentationm results. **(a)**Content images: Hoover Tower, Memorial Church, Gates, and the Chemistry buildings respectively. **(b)** Results from the scikit-image package of Otsu's method. The blue represents the foreground and red the background. **(c)** Results from the OpenCV package of the GrabCut algorithm. The part of the building shown has been classified as foreground. **(d)** Results from SegNet. The red areas have been categorized as "building," light green as "tree," blue as "pavement," pink as "road," and gray as "sky."

E. Foreground extraction using GrabCut

GrabCut, developed by Rother, etal., is another methodology of scene extraction [12]. The GrabCut algorithm uses a Gaussian Mixture Model(GMM) to model the foreground and background. GMM learns a new pixel distribution by labelling unknown pixels as either probable foreground or probable background depending on its relation with other pixels in terms of color statistics, analogous to clustering. A graph is produced from the distribution, where the nodes are pixels, every foreground pixel is connected to a source node, and every background pixel is connected to a sink node. The weights of the edges are defined by pixel similarity and a mincut algorithm is used to segment the graph, continuing until the classification converges.

An fresh feature of GrabCut is its interactivity. The user can specify a bounding box, where the pixels within the box are unknown and those outside are hard- labeled as background. We used the OpenCV package for GrabCut(13) and specified a bounding box roughly surrounding the structure portion of a given content image(Fig. 1(c)).

F. Segmentation using SegNet

SegNet is the deep encoder- decoder architecture formulti-class pixelwise segmentation, created by the Computer Vision and Robotics Group at the University of Cambridge, UK [14]. The model consists of a sequence of encoders,non-linear processing layers, and a corresponding set of decoders followed by a pixelwise classifier. Each encoder is composed of one or further convolutional layers with batch normalization and ReLUnon-linearity, followed by maxpooling andsub-sampling. By maxpooling indicators in decoders to perform

upsampling of low resolution feature charts, the model retains high frequence details in the segmented images. The entire architecture can be trained end- to- end using stochastic gradedescent.

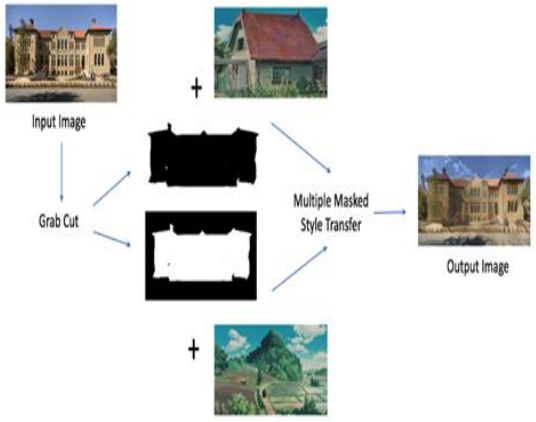


Figure 2. Model Pipeline. After the image is segmented into foreground and background, the best style images are chosen. All are fed into the multiple masked style transfer network to create the output image.

In our case, we used the SegNet demo [15] to produce results(Fig. 1(c)). The model was veritably good at distinguishing the structure, road/ pavement, trees, and sky, which are the main object orders we wanted to use in pixelwise division. still, after comparison with further traditional computer vision ways like GrabCut we decided to use masks from the GrabCut algorithm rather. The main two reasons for this decision were the masks from GrabCut had lower noise, which can be seen as the yellow pixels within the red structures in every sample mask(Fig. 1), and the edges of structures were sharper which is especially apparent in the top roof of the third structure(Fig. 1(iii)). We anticipate to use further nuanced segmentation like SegNet or DeepMask[16] and SharpMask [17] in the future in order to overcome these failings.

4) Finding the Best Frames from Animation

After segmenting the input photograph, we capture two pics , one of the foreground and one of the backgrounds. The rest of the image is filled in with the mean value of the photograph. We also pick an image for the style image for each. For each frame of the film, we calculate the content loss between the frame and the background. We take the frame with the minimal loss as the style image for background. We repeat the process for the foreground. The content loss, also known as the mean squared error(MSE) is defined as

$Loss = \Sigma i,(Pij - Mij)2$

In the end we have two style images, one to be used on the foreground and one on the background.

5) Masked Style Transfer

For style transfer, we modified an existing execution of the neural algorithm of artistic style (19), which uses apre-trained VGG- 16 network. We used two loss functions, one of the content and one of the style. The content loss function taken at one position in the network, and is characterized by the mean square loss of the representations of the input image and the snapshot at those positions. The style loss function is taken at multiple layers in the network, and characterized by the mean squared loss between the Gram matrices of the input image and the artwork. The total loss is also a combination of these two loss functions. We also perform grade ascent on a white noise image to minimize the total loss [2]. We modified this algorithm to be suitable to take in two masks of the content image and two style images, transferring the style widely from each image to its corresponding content image mask.

For both birth and Semantic Segmentation we decided to keep the original color of the content image in order to drop the number of independent variables when comparing results. likewise, across all tests in baseline and Semantic Segmentation we ran 800 duplications using an L- BFGS optimizer with a literacy rate of 4e- 4.

IV. RESULTS

1) Baseline

The images for baseline were suitable to maintain some content from the input image and some style from the style images. still, there were several issues. The structure edges came distorted due to the sharp corners present in the style image. likewise, without masking off the structures from the rest of the image the background started

hallucinating sharp edges, as seen in the sky in the baseline results(Fig. 3(a i)) and(Fig. 3(a iv)).

2) Masked Style Transfer with Handpicked Style Images

The results from Masked Style Transfer with handpicked style images yielded the best results. We can now see a clear distinction between the two styles transferred to each part of the content image. For example, in Fig. 3(b i) and Fig. 3(b iv) we can see that the sky does not have any sharp straight edges; rather the sky is much more semantically similar to the sky in Studio Ghibli animations with puffy and circular white clouds. Moreover, in Fig. 3(b i) we can see that the trees kept more semantic meaning than compared to baseline, where the trees became distorted. The buildings also showed improvement from baseline, with less distortion while still maintaining some animation-like style. However, there was still some distortion present, as seen in Fig. 3(b iv).

3) Masked Style Transfer with Automated Style Images

The results from Masked Style Transfer with opted style images yielded the stylish results. We can now see a clear distinction between the two styles transferred to each part of the content image. For illustration, inFig. 3(b i) andFig. 3(b iv) we can see that the sky doesn't have any sharp straight edges; rather the sky is much further semantically analogous to the sky in Studio Ghibli animations with fluffy and indirect white shadows. also, inFig. 3(b i) we can see that the trees kept more semantic meaning than compared to baseline, where the trees came malformed. The structures also showed enhancement from baseline, with lower deformation while still maintaining some animation- suchlike style. still, there was still some deformation present, as seen inFig. 3(b iv).

Figure 3. Baseline (A) results, Masked Style Transfer with handpicked style images (B), and Masked Style Transfer with automated style images (C) for Hoover Tower (i), Gates (ii), Memorial Church (iii), and Chemistry Buildings (iv).

V. CONCLUSION AND FUTURE WORK

We presented a system of style transfer that uses segmentation of the content image and content-based selection of style images from animation to maintain higher semantic meaning in the yield image. Our intention was to capture the essence of a certain animation workplace, and we primarily planned on detecting, localizing, and transferring style from objects in style images from movies to corresponding objects in the content image. still, we ran into challenges, similar as poor object discovery for animated images and limited orders for objects to begin with. We determined that we'd concentrate on relating structures in the foreground and the rest of the terrain in the background. After testing a variety of different pixelwise segmentation ways, we chose GrabCut, because of it’s lack of noise and relative computational simplicity. In addition, rather of choosing style images arbitrarily, we used a content loss function to opt particular frames from a given animated film; one style image for the foreground and one for the background, chosen because of their semantic similarity to the content image. With the content image, its segmentation masks, and style images as input, we use apre-trained VGG- 16 network with two loss functions as described in [2], one for content and one for style, to perform segmented style transfer. We begin that

this methodology successfully preserves structural information especially in the foreground, while still being suitable to successfully transfer the Studio Ghibli art style of painted texture and colour for objects in both the foreground and background. likewise, we begin added benefits in sharper structure edges with this technique when compared to baseline. Eventually, while we don't suppose we've fully achieved our intention of attaining the " essence " of a given animation, we believe there are still important operations of being suitable to apply style in a semantically-sensitive way as well as applying different styles to different areas or only some areas of an image. In the future, we hope to use further nuanced semantic segmentation ways, similar as DeepMask [16] and SharpMask [17], for more precise object- to- object style transfer to maintain the semantics of a given content image at an indeed higher degree. We also hope to experiment with other more complex content loss functions to elect frames from animations that are semantically matching to the content image more precisely; for illustration, we could consider other features similar as spatial or structural information. We'd also like to be suitable to transfer to further than just two segments, similar as foreground and background, but rather to semantically segmented elements, similar as structure, tree, sky, etc. Eventually, we'd also like to see if our methodology generalizes to other animation workshops beyond Studio Ghibli or even concentrate on a certain "mood" or atmosphere of a particular part of a film and explore what aspects of style must be transferred to evoke that mood.

REFERENCES

[1] Kevin Yang* Jihyeon Lee* Julia Wang*Stanford University kyang6 Stanford University jlee24 Stanford University jwang22

[2] Karpathy, Andrej. "A peek at trends in machinelearning." https://medium.com/@karpathy/a-peek-at-trends-in-machine-learning-ab8a1085a106

[3] L. A. Gatys, A. S. Ecker, and M. Bethge. "A Neural Algorithm of Artistic Style". arXiv: 1508.0657, 2015.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge, A. Hertzmann, E. Shechtman. "Controlling Perceptual Factors in Neural Style Transfer". arXiv: 1611.07865, 2017.

[5] Open Toonz, DWANGO Co. LTD. https://opentoonz.github.io/e/

[6] E. Shelhamer, J. Long, T. Darrell. Fully Convolutional Networks for Semantic Segmentation. arXiv: 1605.06211, 2016.

[7] "Studio Ghibli Creation Process." *The Legacy of Hayao Miyazaki*, 2017.

[8] A. J. Champandard. Semantic Style Transfer and Turning Two-Bit Doodles into Fine Artworks. arXiv: 1603.01768, 2016.

[9] X. Liang, B. Zhuo, P. Li, L. He. CNN based texture synthesize with Semantic segment. arXiv:1605.04731, 2016.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr. Conditional Random Fields as Recurrent Neural Networks. arXiv:1502.03240, 2015.

[11] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". *IEEE Trans. Sys., Man., Cyber*. **9** (1): 62–66.

[12] Otsu thresholding through the scikit-image package. http://www.scipy-lectures.org/packages/scikit-image/#binary-segmentation-foreground-background

[13] C.Rother, V.Kolmogorov, A.Blake, "GrabCut": interactive foreground extraction using iterated graph cuts, ACM Transactions on Graphics (TOG), v.23 n.3. doi:10.1145/1015706.1015720, 2004.

[14] Interactive foreground extraction using GrabCut algorithm through the OpenCV package. http://docs.opencv.org/trunk/d8/d83/tutorial_py_grabcut.html

[15] Vijay Badrinarayanan, Ankur Handa and Roberto Cipolla "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling." arXiv:1505.07293, 2015.

[16] SegNetDemo. http://mi.eng.cam.ac.uk/ projects/segnet/ demo. Php

[17] P. Pinheiro, R. Collobert, P. Dollar. Learning to Segment Object Candidates. arXiv: 1506.06204, 2015.

[18] P. Pinheiro, T. Lin, R. Collobert, P. Dollar. Learning to Refine Object Segments. arXiv: 1603.08695, 2016.

[19] C. Li, M. Wand. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. arXiv: 1601.04589, 2016.

[20] Tensor Flow (PythonAPI)implementation of NeuralStyle. https://github.com/cysmith/neural-style-tf