Improved Style Transfer with Semantic Segmentation for High-Quality Image Stylization

Yin Chen, Fangming Cheng, Owen Zhou
Department of Computer Science, Rice University
yc153@rice.edu, fc31@rice.edu, zz102@rice.edu

Abstract

Style transfer has gained significant attention lately due to its potential in studying fundamental challenges related to image understanding and synthesis. In our project, we utilize semantic segmentation to differentiate an image into different classes and apply style transfer to the individual classes. By including extra content layers in the style transfer model, we improved the performance of the model, resulting in images with enhanced clarity of contours, greater color uniformity, and heightened distinguishability. We evaluated the effectiveness of our approach by captioning both the original and transferred images and comparing the similarity between the image captions using cosine similarity as a metric. As a result, we developed a method that allows us to transfer the style of certain objects or backgrounds of a picture, and output a captivating image.

1. Introduction

Image style transfer is a popular task that can be done using many different methods, previous work that introduces different approaches to complete this task include [1, 2, 3]. Our project aims to beautify images through the mixture of various styles and the incorporation of diverse colors. We introduce a method that integrates semantic segmentation, style transfer, and image captioning, which is a challenging task that demands the combination of multiple complex computer vision techniques. We used the vgg-19 pretrained model as our Convolutional Neural Network and improved its performance by including extra content layers. This leads to a noticeable increase in output image quality. In addition, semantic segmentation is used to separate the image into objects and background, allowing us to perform style transfer on individual components, resulting in a more captivating output image. Evaluation is done by generating captions for both the original image and the style-transferred image. We quantified the effect of style transfer on the semantic components of the image by measuring the cosine similarity of the image captions, which indicate that style transfer significantly impacts the caption similarity between the original image and the output image.

2. Related Work

Early work on image style transfer include [1], which uses a neural algorithm of artistic style that can separate and recombine the image content and the style of the natural images. More recent works like [2], which uses reversible neural flows and an unbiased feature transfer module that support both forward and backward inferences and operate in a projection-transfer-reversion scheme. [3] introduces another style transfer method by proposing a dual-consistency loss to train an encoder-decoder network with adversarial discriminator. Our project refines the style transfer function from the vgg-19 pre-trained model by including extra content layers to improve model performance, which increases output image quality.

Previous work on semantic segmentation include [4, 5, 6], which utilizes the Fully Convolutional Network (FCN) model for image semantic segmentation. However, we choose to use the DeepLabV3 model instead since it is advantageous in capturing more contextual details and segmenting objects with large spatial extent.

Image captioning is used in many applications. Early work such as [7], which proposes a novel deep hierarchical encoder-decoder network for image captioning. More recent work include [8, 9], which utilizes the transformer architecture and develops an image transformer for automatic image captioning. Our project uses the Multi30k dataset as a foundation for image captioning, and we employ a custom model with a ResNet-152-based image encoder and an LSTM-based text decoder. Image captioning is applied to both the input and output image, serving as a tool for evaluation in our experiments.

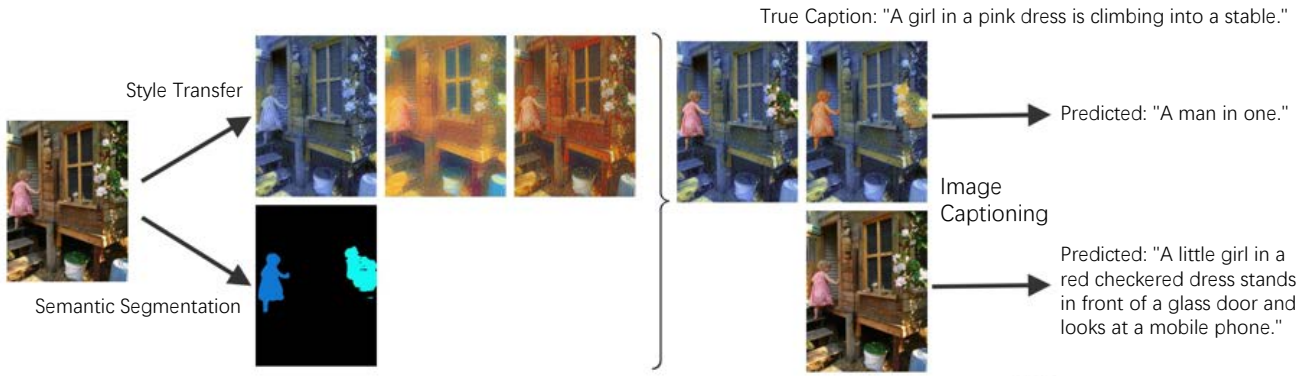

Figure 1. Here we show the process of our project in how to combine style transfer, semantic segmentation, and image captioning.

3. Model

3.1. Style Transfer Model

We used the vgg-19 pre-trained model as our Convolutional Neural Network, which is highly accurate and easy for transfer learning. We also made a new Convolutional Neural Network to put in modules that are supposed to be activated sequentially. In addition, we wrote a function that creates a new neural network model that is a modified version of our pre-trained convolutional neural network(vgg-19). What's more, we added normalization as the first layer in this model so the input data will be normalized before being fed into the subsequent layers, which improves the stability and convergence speed. This modified CNN is used to generate a new image that combines the content of one image and the style of another image.

3.2. Semantic Segmentation Model

We used DeepLabV3 as our semantic segmentation model, which is accurate and efficient. It uses the ResNet models as the backbone along with Atrous Convolution and Atrous Spatial Pyramid Pooling model. A model similar to DeepLabV3 is the Fully Convolutional Network(FCN) model, which is composed of only convolutional and pooling layers. The advantage of using DeepLabV3 over FCNs is its ability to capture more contextual information, which is useful for segmenting objects with large spatial extent. Therefore we chose to use DeepLabV3 as our semantic segmentation model.

3.3. Image Captioning Model

The Multi30k dataset serves as the foundation for our image captioning experiment, employing a custom model with a ResNet-152-based image encoder and an LSTM-based text decoder. The image captioning component takes two distinct input types: the original image and the image subjected to style transfer and segmentation.

The image encoder is responsible for extracting essential visual features from the input image. Here we used ResNet-152 as the convolutional neural network for capturing complex and hierarchical visual patterns. The LSTM-based text decoder takes the extracted features and generates a caption that describes the image content.

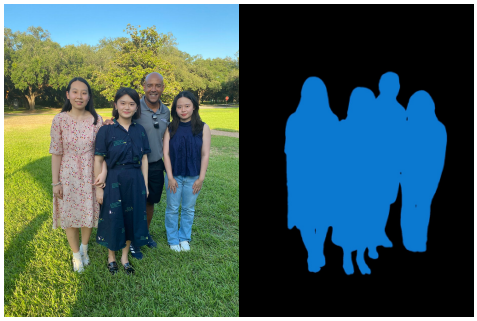

Figure 2. Here's an example of how Semantic Segmentation can be used to segment people from the background.

In our project, the image captioning model is applied to both the original image and the image after style transfer and segmentation. The generated captions are then compared to the true captions using cosine similarity as a metric. By examining the differences in the cosine similarity scores, we can evaluate how style transfer and segmentation affect the preservation of content and semantic information in the original image.

4. Experiments and Results

4.1. Combine Style Transfer and Semantic Segmentation

As shown in Figure 3 and Figure 4, we can transfer the style of one image into another image. By combining it with semantic segmentation, we can identify the human and object classes present in the image, and select the style that

Figure 3. The style of the target image will be transferred to the style of each of these images.

Figure 4. Here are three outputs generated by the style transfer model, which have the same styles as the images in figure 3.

each individual class will be transferred to.

4.2. Style Transfer Improvements

Upon refining the style transfer function, we included conv_5, conv_6, and conv_7 as extra content layers in the style transfer model, resulting in improved model performance. This adjustment enables a more comprehensive extraction of content features, capturing diverse levels of details in the input image. The resulting images exhibit enhanced clarity of contours, greater color uniformity, and heightened distinguishability. We have conducted experiments and included contrast images in this section to demonstrate the improvements. By selecting a target image and a style image used to transfer the style of the target image, we generated two outputs using the original style transfer model and the improved model and made a comparison.

As shown in Figure 5, we generated two style-transferred images using the original style transfer model and the improved model. The one on the bottom left is generated by the original model, and the one on the bottom right is generated by the improved model. By contrasting these two images, we can see that the bottom right has a clearer contour with even color, while the bottom left image is hard to identify and the color is uneven. By using the improved style transfer model, we can significantly enhance the visual appeal of the input image.

4.3. Cosine Similarity

We investigated the effect of style transfer on image captioning using cosine similarity as a metric. The results indicate that the original images have a higher cosine similarity with the real captions compared to the output images.

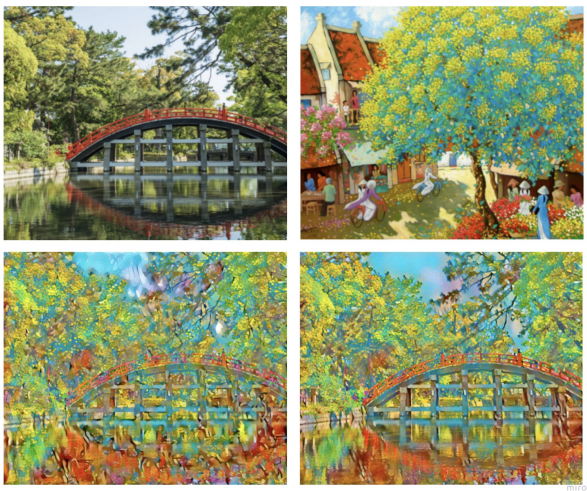

Figure 5. Here is an example shows the effect of using the improved style transfer model.

Cosine similarity is a metric for measuring the similarity between two vectors, in this case, the TF-IDF vectors of the captions. By calculating the cosine similarity between the captions of the original images and the style-transferred images, we can quantify the degree to which the style transfer affected the semantic content of the captions.

As shown in Figure 6 and Table 1, our evaluation results demonstrate that style transfer has a significant impact on the similarity between the original image captions and the transferred image captions. In Figure 6, the x-axis represents the indexes of the 25 images input we used, and the y-axis represents the cosine similarity(tf-idf) between captions. The orange bar shows the cosine similarity between the true caption and the predicted caption of original images. The green bar represents the cosine similarity between the true caption and the predicted caption of output images, which have been segmented, as well as all the objects and backgrounds of which had style transfers applied using different style images. The blue bar shows the cosine similarity between the original images and the output images. In Table 1, we calculated the average and maximum cosine similarity for these three situations, concluding that the cosine similarity decreased by 52.06% after style transfer and segmentation. As a result, the similarity between the captions of the original image and the transferred image is greatly affected by style transfer.

5. Conclusion

The objective of our project is to devise an optimal approach for enhancing the visual appeal of images using different techniques. By using semantic segmentation, we differentiate the input image into separate classes and apply style transfer on individual classes to generate output im-

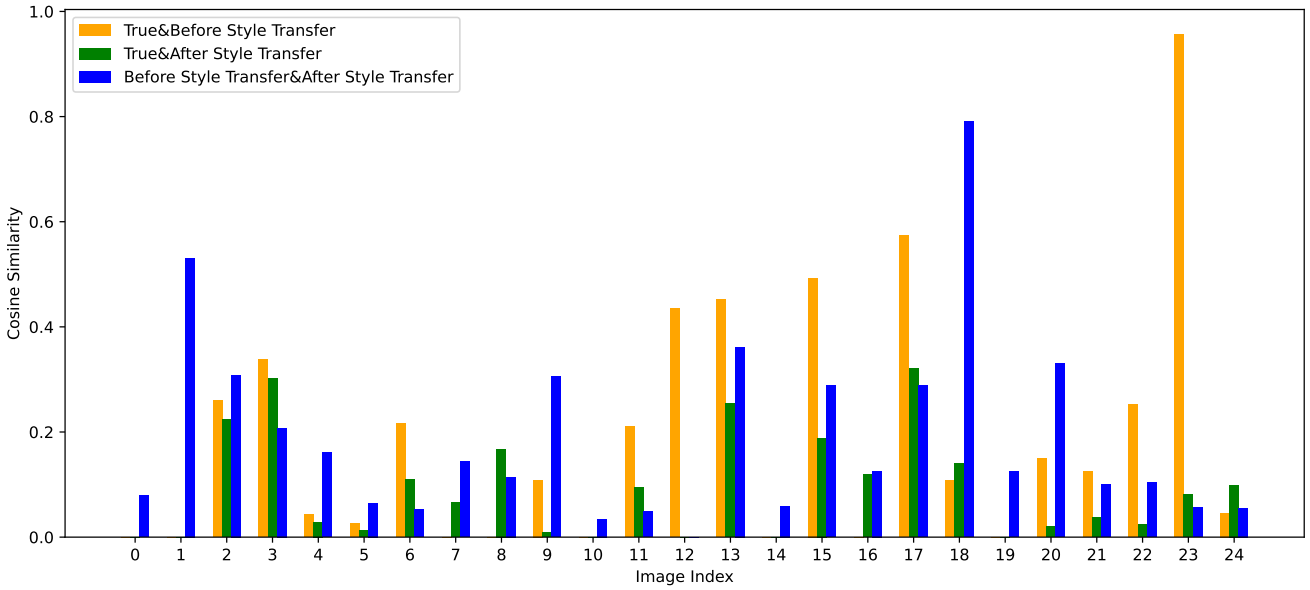

Figure 6. Cosine similarity

Types	Average Cosine Similarity	Maximum Cosine Similarity
True + Original	0.1917	0.9559
True + Styled and Segmented	0.0919	0.3208
Original + Styled and Segmented	0.1895	0.7910

Table 1. Average and Maximum Cosine Similarity Results

ages that are more visually appealing. In addition, we modified the style transfer model by including extra content layers, which increased the image quality of the output. By using cosine similarity as a metric, we show that style transfer significantly impacts the similarity between the captions of the output image and the original image. In conclusion, our project is capable of producing visually captivating output images of higher quality while preserving additional features from the input image.

6. Appendix

For code, please click here.
For slides please click here.
Our demo video is also uploaded here.

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

[2] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "Artflow: Unbiased image style transfer via reversible neural flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 862–871, 2021.

[3] Z. Ma, J. Li, N. Wang, and X. Gao, "Semantic-related image style transfer with dual-consistency loss," *Neurocomputing (Amsterdam)*, vol. 406, pp. 135–149, 2020.

[4] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-fcn for image semantic segmentation," in *Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks, ISNN 2019, Moscow, Russia, July 10–12, 2019, Proceedings, Part I 16*, pp. 97–105, Springer, 2019.

[5] H. Zhou, J. Zhang, J. Lei, S. Li, and D. Tu, "Image semantic segmentation based on fcn-crf model," in *2016 International Conference on Image, Vision and Computing (ICIVC)*, pp. 9–14, 2016.

[6] S. Kumar, A. Negi, J. Singh, and H. Verma, "A deep learning for brain tumor mri images semantic segmentation using fcn," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–4, 2018.

[7] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder-decoder network for image captioning," *IEEE transactions on multimedia*, vol. 21, no. 11, pp. 2942–2956, 2019.

[8] S. He, W. Liao, H. R. Tavakoli, M. Y. Yang, B. Rosenhahn, N. Pugeault, H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, "Image captioning through image transformer," in *Computer Vision - ACCV 2020*, vol. 12625 of *Lecture Notes in Computer Science*, pp. 153–169, Switzerland: Springer, 2021.

[9] X. Li, W. Zhang, X. Sun, and X. Gao, "Semantic-meshed and content-guided transformer for image captioning," *IET computer vision*, vol. 16, no. 5, pp. 431–444, 2022.