Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/73/20240403

Content-Adaptive Style Transfer

Kaifeng Teng

Donghua University, 2999 North Renmin Road, Songjiang District, Shanghai, 201620, China

lawrenceteng0125@gmail.com

Abstract. The development of Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN) has catalysed the emergence of the Neural Style Transfer (NST) field. NST elucidates the computer's comprehension of artistic elements, holding paramount significance in both academic and practical domains. In this paper, we introduce a novel system termed "Content-Adaptive Style Transfer." This system incorporates two additional techniques: object segmentation and image fusion, augmenting the customizability of NST. Our experimental results underscore its capability to execute style transfer exclusively on the foreground, leaving the background unaffected, or to modify the background style while preserving the original foreground. These innovations promise more rational outcomes in specific scenarios.

Keywords: Generative Adversarial Network, Convolutional Neural Network, Neural Style Transfer, Object Segmentation, Image Fusion.

1. Introduction

The field of computer vision is currently experiencing rapid development and evolution. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has led to significant advancements in areas such as image classification, object detection, and facial recognition [1]. Furthermore, Generative Adversarial Networks (GANs) have emerged as a pivotal and innovative technology. Introduced in 2014, GANs represent a type of generative model that has yielded remarkable applications, delivering impressive results in both performance and quality [2].

Within the broad spectrum of GAN applications in image processing, they can be predominantly categorized into two types: image synthesis and image-to-image translation [3]. Image synthesis entails the creation of new images from an existing dataset, a text description, or additional input data. For instance, GANs can generate new handwritten digit images by learning from the MNIST dataset. They can also generate multiple plausible images matching a given text description. Image-to-image translation, on the other hand, aims to learn a mapping between an input image and an output image. It can separate the style and content of an input image, allowing the replacement of the original style with a new one, transforming the input image's style, such as converting a horse into a zebra (see Figure 1) or a summer landscape into a winter scene [4]. Furthermore, image-to-image translation encompasses data augmentation and image restoration as distinct applications.

In the context of image synthesis, conditional GAN stands as a fundamental model [5]. It is trained

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/73/20240403

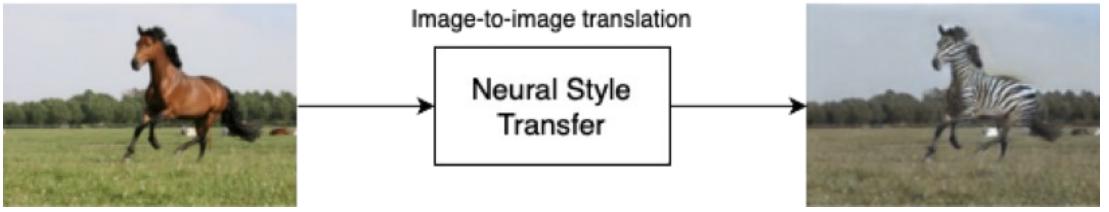

Figure 1. Style transfer transforms the horse in the original image into a zebra.

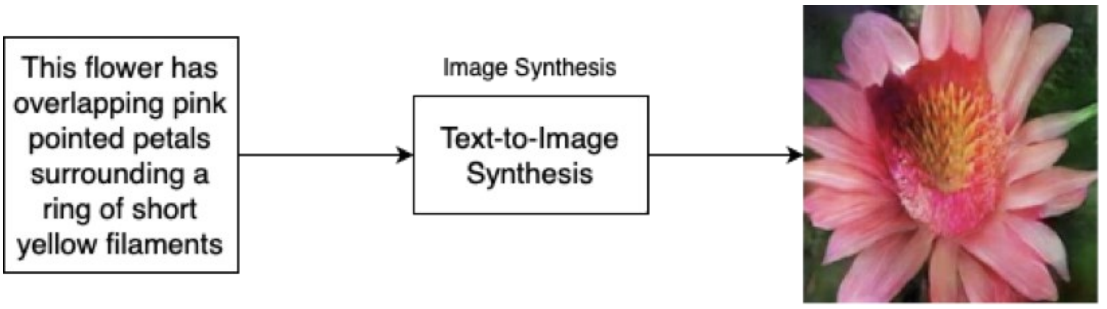

Figure 2. Text-to-Image synthesis accepts a text description as input and outputs an image that matches the description.

on labelled datasets, allowing for the specification of labels for each generated instance. While an unconditional MNIST GAN generates random handwritten digit images, a conditional MNIST GAN permits the specification of the digit image to be generated. LAPGAN extends the conditional GAN by enabling both the generator (G) and discriminator (D) networks to accept additional label information as input, a common practice for improving image quality [6]. Subsequently, the integration of natural language into this conditional framework has been achieved. For instance, a trained GAN can generate multiple coherent images matching a text description, such as "a black dog with white spots, a long tail, and drooping ears" (see Figure 2). The Generative What-Where Network (GAWWN) [7] adds the image's location as a condition to text descriptions, offering an interactive interface for progressively constructing larger images using text descriptions and provided boundaries.

In the realm of image-to-image translation, Conditional GAN serves as the foundation for various other models. The pix2pix model, based on CGAN, provides a general solution for tasks that translate input images into output images [8]. It is not constrained to a specific application and can be applied to a wide range of tasks, such as colorizing black and white photos, converting Google Maps images into aerial views, or transforming sketches into photos. Generally, training a model for image-to-image translation necessitates a dataset containing paired examples. For instance, when translating a summer landscape into a winter scene, the dataset comprises input images (summer landscapes) and corresponding modified versions serving as the expected output images (winter landscapes). However, the requirement for paired training datasets poses significant limitations, as preparing suitable datasets can be challenging or even impossible in certain cases, such as with famous paintings and their corresponding photographs. Therefore, there is a need for a model that can perform unpaired image-to-image translation. CycleGAN fulfils this need and has proven successful for unpaired image-to-image translation [9]. It extends the GAN architecture by simultaneously training both G and D networks. One generator transforms input images from the first domain into images resembling the second domain, while the other generator performs the reverse transformation. The discriminator classifies the generated images and updates the generators. Nevertheless, this extension alone is insufficient for generating translations of input images. To address this, the concept of cycle consistency is introduced. Inspired by machine translation, where a piece of English text translated into French and then back into English should match the original text, CycleGAN adopts a similar concept. The first generator transforms an

DOI: 10.54254/2755-2721/73/20240403

input summer photo into a corresponding winter photo. The generated result is then used as input for the second generator, which transforms the winter photo back into a summer scene, closely resembling the original input. A loss function is employed to measure the similarity between the output of a translation, the reverse translation cycle, and the original photo. By continuously optimizing this loss function, unpaired image-to-image translation is achieved.

The increasing maturity of these techniques has enabled computers to perceive artworks to a certain extent, leading to the emergence of Neural Style Transfer (NST). NST employs CNNs to render a content image in various artistic styles. Previously, this process demanded skilled artists to invest significant time, whereas now it can be swiftly accomplished by computers. NST holds significant academic value. Art was traditionally a distinctly human domain, demanding emotions, imagination, and creativity. This relates to human emotional cognition, a challenge for AI to replicate. However, NST has made substantial strides in this regard. Many NST implementations have partially learned the styles of artists like Monet and Van Gogh, bridging the gap between technology and traditional art, thereby highlighting AI's extraordinary potential in the realm of art. Moreover, NST boasts numerous practical applications, aiding artists in creative endeavours, contributing to innovative film production, and even serving as an entertainment product integrated into daily life [10].

In this work, a higher degree of customization in style transfer is pursued. To achieve this objective, a novel system is designed that not only incorporates GANs but also integrates object segmentation and image fusion. Although CycleGAN has already enabled style transfer to produce realistic and impressive results, one challenge remains. In an image, certain elements or objects cannot be individually transformed while keeping the rest unchanged. This necessitates the separation and recognition of image elements, introducing the concept of object segmentation [11]. Following this process, numerous segmented image elements are obtained, which need to be reassembled into a complete image as the output. To meet this requirement, image fusion is introduced [12].

In conclusion, this paper builds upon the research of predecessors to integrate three techniques from the field of computer vision: GANs, object segmentation, and image fusion. This integration achieves a higher degree of customization in style transfer.

2. Pix2pix

Pix2pix is a CGAN-based model that offers a versatile solution for various image-to-image translation tasks, including colorizing black and white photos and transforming sketches into photographs. Unlike traditional CGANs, pix2pix introduces some modifications to both the G and D networks.

In the G network, pix2pix employs a U-Net architecture initially used for the segmentation of biomedical images [13]. This architecture involves an encoder-decoder network, where the input undergoes downsampling through multiple layers and reaches a bottleneck layer. Subsequently, the process is reversed to produce the output, allowing information to pass through all layers, including the bottleneck. For many image-to-image translation tasks, there is shared low-level information between the input and output, so U-Net introduces skip connections between encoders and decoders (see Figure 3). These skip connections enable the generator to bypass the bottleneck layer and transmit this valuable information.

DOI: 10.54254/2755-2721/73/20240403

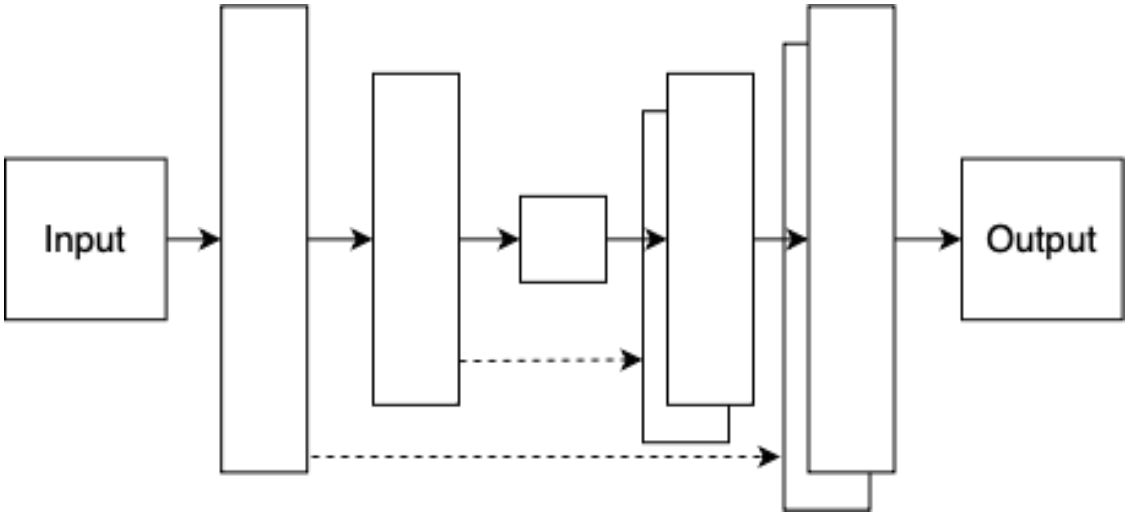

Figure 3. U-Net introduces skip connections between encoders and decoders, allowing the generator to bypass the bottleneck layer.

or the D network, pix2pix introduces PatchGAN. In contrast to conventional GAN discriminators that take the entire image as input and classify it as real or fake, the PatchGAN discriminator classifies all N × N-sized patches within the image and averages the results to obtain the final outcome. The patch size N is a hyper-parameter that requires tuning for the discriminator, and experiments have demonstrated that N=70 yields superior results. This approach offers faster execution, a reduction in the number of parameters, and the ability to produce sharp high-frequency details.

Additionally, pix2pix incorporates L1 distance into the loss function. While the original GAN employs binary cross-entropy as the loss function to distinguish real from fake images, many image-to-image translation tasks demand not only distinguishing real from fake but also measuring how closely the generated image resembles the target image. L1 distance represents the mean absolute pixel value differences between two images, with smaller values indicating greater similarity. When two images are identical, the L1 distance is 0.

3. Style Transfer for Foreground Object

This paper introduces a system designed to achieve a higher degree of customization in style transfer, consisting of three main components: object segmentation, style transfer, and image fusion. While previous work in the Neural Style Transfer (NST) field has primarily focused on optimizing style transfer's performance and quality, a persistent issue remains: the inability to transform a specific element in an image while maintaining the rest of the image unchanged.

For example, in the transformation of a classic horse into a zebra, the horse in the original image can be considered the foreground, while the surrounding grass and trees can be seen as the background. During the style transfer process, the horse in the foreground transforms into a zebra, but it also alters the colours of the grass and trees in the background, resulting in an image with a different background style. While this overall transformation is suitable for some scenarios, such as when turning a real-world photograph into an artistic work with a particular style, there are cases where only the foreground should change—like turning the horse into a zebra while preserving the background's original colour tones.

To address this issue, object segmentation is introduced to separate the foreground from the background, enabling the application of style transfer exclusively to either the foreground or the background. Alternatively, it allows for transforming the foreground into one style and the background into another style, thereby enhancing style transfer's customizability and expanding its potential.

Following the separate processing of the foreground and background, image fusion is employed to recombine these two parts into a single image, producing the final result. In summary, our system's

DOI: 10.54254/2755-2721/73/20240403

processing workflow involves using object segmentation (e.g., Segment Anything [14]) to separate the image into its foreground and background. Subsequently, style transfer (utilizing CycleGAN or Pix2pix) is applied independently to the foreground and background. Finally, image fusion is used to combine these two processed elements into a single image, resulting in the final output (see Figure 4).

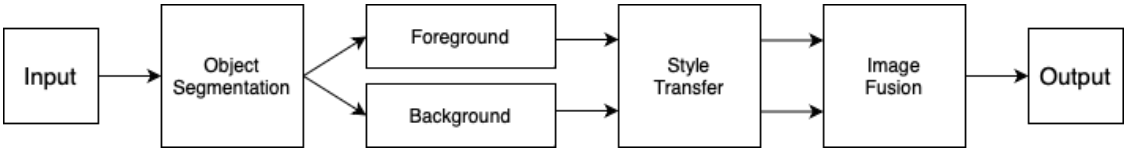

Figure 4. After applying object segmentation to the input image, the foreground and background are extracted. Style transfer is then applied individually to the foreground and background, and the final result is generated by fusing these two processed parts using image fusion.

4. Experiment Results

To validate the viability of the previously outlined system, several classic applications in the field of style transfer were chosen: converting a horse into a zebra, transforming an apple into an orange, and changing a summer scene at Yosemite National Park into a winter one. Experiments were carried out for these tasks, and an analysis of the system's capabilities and limitations was conducted based on the outcomes.

In the transformation of a horse into a zebra (see Figure 5 and 6), the system effectively achieved its intended objective by converting the horse, representing the foreground, into a zebra while leaving the original image's background untouched. Notably, the colours of the trees and grass in the background of the resulting image remained unchanged, thus addressing the limitation of traditional style transfer methods that transform the entire image as a whole without the ability to target specific components.



Figure 5. The left image depicts the original picture, the middle image portrays the foreground transformed through style transfer, and the right image shows the resulting output.

DOI: 10.54254/2755-2721/73/20240403



Figure 6. The left image depicts the original picture, the middle image portrays the foreground transformed using style transfer, and the right image represents the final result.

In the transformation of an apple into an orange (see Figure 7), the style transfer process led to a darkening of most of the background, resulting in a mismatch with the original lighting conditions. However, our system successfully restored the original lighting effects, enhancing the overall realism of the outcome. The original image had a reddish tint in the apple's shadow, implying that the shadow of the orange in the result image should have had some orange hue. Unfortunately, this aspect was not accurately addressed, causing a lack of coherence in the final image. This example reveals a limitation of our system: when only the foreground of an image is transformed while the background remains unchanged, elements in the background related to the foreground, such as shadows, do not align with the transformed foreground, thereby impacting the quality of the result.



Figure 7. The left image illustrates the original picture, the middle image displays the original image transformed through style transfer, and the right image shows the final output.

In the previous examples, the emphasis was on transforming the foreground while preserving the background. However, there are cases where transforming the background while preserving the foreground is desirable. In the transformation of a summer scene at Yosemite National Park into a winter one (see Figure 8 and 9), an image featuring a cabin in the woods was chosen. Style transfer effectively altered the background to a winter scene, but it also needlessly changed the roof of the cabin from its original deep red colour to deep grey. This was an unnecessary alteration. Figure 9 provides a similar example where style transfer transformed mountains, forests, and a stream, but it also turned the person's skin colour into an extremely unrealistic light blue. In such scenarios, preserving the foreground and only modifying the background would result in more favourable outcomes.

DOI: 10.54254/2755-2721/73/20240403



Figure 8. The left image portrays the original picture, the middle image depicts the original image transformed through style transfer, and the right image displays the final result.



Figure 9. The left image represents the original picture, the middle image portrays the original image transformed using style transfer, and the right image displays the final output.

Furthermore, this system can also function as an auxiliary tool for style transfer to correct elements that appear unrealistic after transformation (see Figure 10). For instance, in Figure 10, the colour of the sky in the transformed image significantly deviates from the original, creating an unrealistic appearance. Addressing this issue solely through style transfer might necessitate a larger dataset and more complex networks. In our system, with the assistance of object segmentation, the original sky can be directly applied from the original image to the result image. Although this step may not be entirely rigorous, it can rapidly yield higher-quality results.

5. Conclusion

The results outlined in this paper vividly showcase the diverse application scenarios of content-adaptive style transfer. Beyond the capacity to independently manipulate the background and foreground while making informed decisions regarding style adjustments, the introduction of object segmentation empowers us to address each element within an image individually. This breakthrough ushers in the potential to impart multiple styles to images featuring various elements. When coupled with well-suited images and a discerning artistic sensibility, this approach can yield profoundly innovative and awe-inspiring effects.

References

[1] Voulodimos, Athanasios, et al. "Deep learning for computer vision: A brief review." Computational intelligence and neuroscience 2018 (2018).

DOI: 10.54254/2755-2721/73/20240403

[2] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).
[3] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, "Generative Adversarial Networks: An Overview," in IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53-65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.
[4] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." arXiv preprint arXiv:1411.1784 (2014).
[6] Denton, Emily L., Soumith Chintala, and Rob Fergus. "Deep generative image models using a laplacian pyramid of adversarial networks." Advances in neural information processing systems 28 (2015).
[7] Reed, Scott E., et al. "Learning what and where to draw." Advances in neural information processing systems 29 (2016).
[8] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[9] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.
[10] Jing, Yongcheng, et al. "Neural style transfer: A review." IEEE transactions on visualization and computer graphics 26.11 (2019): 3365-3385.
[11] Li, Yin, et al. "The secrets of salient object segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[12] Zhijun Wang, D. Ziou, C. Armenakis, D. Li and Qingquan Li, "A comparative analysis of image fusion methods," in IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 6, pp. 1391-1402, June 2005, doi: 10.1109/TGRS.2005.846874.
[13] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.
[14] Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).